Selecting and Evaluating Models to Reflect Underlying Scientific Principles: Using Basis Sets to Parameterize Hypotheses

by

Karen Emily Nielsen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2017

Doctoral Committee:

Professor Richard D. Gonzalez, Chair
Professor Pamela Davis-Kean
Dr. Brenda K. Gunderson
Assistant Professor Shuheng Zhou

Karen Emily Nielsen

karenen@umich.edu

ORCID iD: 0000-0003-3771-5272

ACKNOWLEDGEMENTS

I would like to express my appreciation to my co-authors. My gratitude goes to my advisor, Dr. Richard Gonzalez, for his guidance and support during the work in this dissertation and other projects. I could never quantify the impact he has had on my scholarly development—he has provided me with immeasurable opportunities, feedback, and encouragement. Thanks also to Dr. Brenda Gunderson, with whom I have greatly enjoyed collaborating on Chapter IV of this dissertation. It was from her that I learned that education will always continue to evolve, even if the context seems static.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES v

LIST OF TABLES vi

ABSTRACT vii

CHAPTER

I. Introduction 1

II. Comparison of Common Metrics in Event-Related Potential Analysis 4

2.1 Introduction 4
2.2 The ERP Signal and Analysis 6
2.3 Comparison of Maximum and Average: An Example 9
2.4 Distributional Theory of Summary Metrics 15
2.5 Simulating Data from an ERP Component 21
2.5.1 Confirmation of Distributional Results 25
2.5.2 Unbalanced Designs via Trial Counts 28
2.5.3 Sampling Rates 31
2.5.4 Latency Differences Across Window Locations and Widths 33
2.6 Discussion and Future Work 37

III. Basis Sets for Testing Meaningful Landmarks in Time Series Data 42

3.1 Introduction 42
3.2 Basis Sets 45
3.2.1 Least Squares Review and Basis Sets 45

3.3 Interpretability of Coefficients 47
3.3.1 Reparameterizing the Polynomial 48
3.3.2 Using Different Basis Sets 50
3.4 An Event-Related Potential Example 52
3.5 Regression Spline Mixed Models 57
3.5.1 ERP Combination Example 58
3.6 Conclusions 62

IV. Applet-Based Training for Identifying Appropriate Statistical Methods 65

4.1 Introduction 65
4.2 Applet Development 67
4.2.1 Providing Feedback for All Responses 69
4.2.2 Summary of Performance 69
4.2.3 Repeated Practice 71
4.2.4 Portable, On-the-Go Practice 71
4.3 Assessing the Efficacy of the Learning Tool 72
4.4 Relevance to Other Chapters in Dissertation 76
4.5 Conclusion 81

V. Conclusion 83

APPENDIX 88

BIBLIOGRAPHY 93

LIST OF FIGURES

Figure

2.1	ERP waveform with labeled components	5
2.2	Grand average waveforms comparing simulated P300s	10
2.3	Waveforms comparing two conditions within a single subject	12
2.4	Distribution of maximum compared to Gumbel distribution	26
2.5	Distribution of maximum compared to normal distribution	27
2.6	The distribution of the differences in maxima is approximately normal	27
2.7	Type II errors in unbalanced designs	28
2.8	Type I errors in unbalanced designs	29
2.9	Prototypical components across differing trial counts	30
2.10	Error rates in unbalanced designs with reduced subject variability	31
2.11	Type I error rates as sampling rate varies	32
2.12	Type II error rates as sampling rate varies	32
2.13	Error rates as sampling rate varies, low latency variability	33
2.14	Type I error rates as window size and location vary	34
2.15	Type I error rates as window centering varies, for subjects with low latency variability	34
2.16	Type I error rates as window centering varies, for subjects with high latency variability	35
2.17	Type I Error rates as window width varies	36
2.18	Type II Error rates as window location varies	37
3.1	Fitting two different data transformations	52
3.2	Plot of all trial-level waveforms	55
3.3	Comparison of two conditions using normal kernels with error bars	56
3.4	ERP data for 64 channels of 1 trial	60
3.5	Mixed-Effects model using normal kernels on ERP data	60
4.1	The landing page of the applet	68
4.2	Feedback is given for both correct and incorrect responses	69
4.3	An example results page in NTS	70
4.4	Change in average performance for matched scenarios	75
4.5	Performance based on usage behaviors	77
4.6	One-parameter models of assessment items	79
4.7	Three-parameter models of assessment items	80

LIST OF TABLES

Table

2.1	Comparison of maximum and average from grand-averaged waveforms	10
2.2	Comparison of t-tests for maxima and averages across conditions . .	11
2.3	Comparison of paired t-tests as subject sample size varies	13
4.1	Self-reported gender of students	73
4.2	Class rank of students	73
4.3	Regression results for post-test score (out of 8)	76
A.1	Questions from assessments	92

ABSTRACT

The problem of selecting an appropriate representation for a given dataset is a critical first step in the analysis process. By making use of a particular model, the researcher places an often-unstated set of assumptions on the shape, or functional form, of the data. Sometimes the chosen model and its assumptions may lead to incorrect conclusions, or not even answer the underlying research question. This dissertation explores ways in which model specification is done in practice, the effects it has, and what we can do to address problems with current approaches.

Time series data, particularly biological, are becoming increasingly common as we explore the relationship between biology and behavior. Event-Related Potentials (ERPs), which are brain responses to time-locked stimuli measured using Electroencephalography (EEG), are one example of such data. The goal in ERP research is to make inferences about neural circuits and mechanisms used when responding to stimuli. We first discuss a methodological divide in this context that leads to both interpretation differences and differences in the underlying distributional theory for testing. Through both analytic work and simulation study, we explore the properties of two competing metrics for ERP component amplitude. This study leads to a suggestion that treating the data-generating model as an analysis framework could provide a major step toward a unifying framework that facilitates reproducible research.

Our framework can draw from the substantive expectations researchers have about the shape of individuals' waveforms, particularly in local regions of interest, and trans-

late these verbalized assumptions into mathematical basis sets. These assumptions allow us to derive properties of the representative waveform and implement them as parameters of a statistical model. We then test hypotheses on landmark parameters of the basis sets via multilevel modeling, which allows us to account for temporal patterns, patterns across channels, individual differences, and differences across experimental conditions.

Biological contexts are not the only areas for applications of this basis set approach. Using an example from statistics education, we show that Item Characteristic Curves (ICC) from Item Response Theory (IRT) can also be conceptualized as basis sets, with interpretable parameters that are reflected in the shape of the resulting curve. This context also provides a venue for shifting the current paradigm of using established models without considering the underlying assumptions that they represent—introductory statistics courses. This work is incorporated into a more general paper on statistics education where we contrast a traditional method of analysis with the basis set approach presented here. We believe that by emphasizing the process of selecting a correct model early in methodological training, we can encourage scientists to be receptive of models that test their hypotheses directly and to begin incorporating the process of creating these models into their standard practice.

Overall, the collection of three papers assembled in this dissertation makes the following contributions: identification of a methodological divide in ERP research and exploration of the properties of two competing metrics, description and demonstration of how basis sets can be designed with meaningful landmarks as parameters, highlighting of additional uses for such basis sets, and emphasis on the value of methodological training for interdisciplinary awareness of the importance of model selection.

CHAPTER I

Introduction

Too often, we apply an analytic model to a problem without considering the properties of the model itself. Selecting the appropriate representation for data can be a powerful step in the analytic pipeline. A model can be designed to address the underling science directly, rather than trying to pigeonhole a research question into a common statistical framework. Throughout this dissertation, basis sets will be used as a way to incorporate landmark points into the parameterization of the analysis model. This facilitates direct testing of research hypotheses and can suggest areas for future studies, making the approach broadly beneficial.

Chapter II explores a methodological divide in the context of neuroscience. This divide leads to both interpretation differences and well as differences in the underlying distributional theory. By choosing to use either the maximum or average as the summary metric for Event-Related Potential (ERP) waveforms, researchers are accepting a set of properties related to the metric. This paper reveals that as features of the data-generating mechanism change, error rates in standard testing procedures change in systematic ways. The paper discusses these findings as they related to the choice of summary metric and ultimately concludes with a suggestion that treating the data-generating model as an analysis framework could provide a major step towards a unifying framework which facilitates reproducible research. This approach

to analysis is explored in greater detail in the following chapter.

Chapter III starts with an introduction to basis sets, followed by a simple example that shows that reparameterization can lead to mathematically equivalent but pragmatically advantageous models. Thus, reparameterizing may yield more interpretable results or results which address scientific theory more directly. However, there are many ways to design a model around meaningful landmarks, and often these models will not be mathematically equivalent. This raises the question of how to design a problem-appropriate basis set. The solution comes from Chapter II: select a basis set that reflects known underlying science and allows for testing of hypothesized properties of the data such that there is a direct link between the research hypotheses and the coefficients of the model. The remainder of the paper expands on the proposed ERP analysis model and demonstrates how to use it for both confirmatory and exploratory data analysis.

The final paper, Chapter IV, approaches the idea of selecting an appropriate testing framework from a new angle—statistics education. Since team science is becoming increasingly interdisciplinary, it is valuable for everyone to be taught about the importance of using models that reflect the underlying science and can be leveraged to test hypothesized theories. One way in which statisticians can promote this model mindfulness is by aligning introductory courses with the goal of reflecting the science in the testing framework. For example, the model from the previous two chapters allows for testing of the peak amplitudes of ERP data directly, while simultaneously controlling for potentially confounding latency differences. This chapter describes an applet for providing students with productive practice of this skill and assesses the efficacy of this tool. In the introductory statistics context this skill is operationalized in terms of selecting an appropriate hypothesis test from standard curriculum. The chapter also relates the idea of generalizing a model framework as a basis set with meaningful parameters via a complementary analyses of the statistics education data

presented in the chapter. The chapter illustrates how the item characteristic curves of item response theory can be reimagined as basis set functions and discusses the relevance of this data analytic approach to the broad field of testing and assessment.

The conclusion summarizes these contributions to the field of statistics. The specific research questions, results, and possible limitations of each paper are summarized and integrated. The overall learnings are then discussed, and extensions are proposed.

CHAPTER II

Comparison of Common Metrics in Event-Related Potential Analysis

2.1 Introduction

Event-Related Potential (ERP) studies are a prominent fixture in the psychophysiological literature. An ERP is a brain response to a time-locked stimulus or response. These are commonly measured using the scalp electrodes of electroencephalography (EEG) to capture voltage fluctuations. ERPs are typically recorded within the first second after stimulus presentation, but alternative time windows relative to either stimulus or response onset are sometimes used. The series of positive and negative voltage deflections that compose the ERP waveform indicate underlying components, some of which are well-studied. Figure 2.1 shows the first 500 milliseconds of a prototypical ERP waveform with several components labeled. Analysis of ERP data usually focuses on the amplitude of each component. However, researchers use different methods for quantifying and assessing amplitude.

There are two summary metrics that are commonly used in the ERP literature to quantify the amplitude of a component. One approach is to take the average of all amplitudes in a prespecified window around a hypothesized component, while the other is to take the peak, or maximum, amplitude within a window. Many

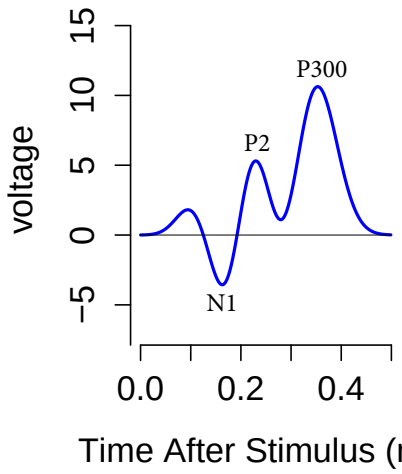


Figure 2.1: ERP waveform with labeled components

ERP analysis guides [10, 17, 19, 20, 37, 41, 42, 46, 48] provide best practices for both summary metrics. Occasionally, these recommendations are discussed in terms of potential impacts on analysis, but without going into specific detail. To date, no reports have explored the extent of these impacts and discussed in depth the rationale for them, nor has a strong statement been made in favor of one or the other.

There is currently not an established standard within the larger body of ERP researchers publishing in major journals, or the editors of these journals. Of the 49 articles using ERP data published in the journal *Psychophysiology* in 2014 (some of which contained multiple studies and may be counted more than once here), 37 use the average amplitude in a prespecified window, 11 use the maximum amplitude in the window, and 5 use a combination (such as the average near a local maximum).

Extreme value theory, a branch of statistics dealing with order statistics and extreme values, suggests that we should be concerned about the use of a local maximum in a context where the average is otherwise appropriate. The maximum has been well-studied in the statistics literature [see, for example, 14] and is known to have different distributional properties from the average.

In this paper, we will outline where these properties might impact analyses, review

relevant assumptions, and assess the appropriateness and potential impacts of using maxima in these analyses. We will begin by describing how ERP data is typically analyzed by providing a simple step-by-step framework. Then, we give an illustrative example of a case when test results for the maximum and average do not agree. This example will serve as the motivation for the remaining sections, in which we explore distributional theory and use simulations to determine how related decisions interact with the summary measure to impact results. We conclude with an overview of findings and ideas for how existing statistical methodology may be introduced to ERP research to circumvent some of the issues uncovered in this paper.

2.2 The ERP Signal and Analysis

First, it is helpful to know a little about how ERPs are traditionally generated, collected, and analyzed. ERPs are considered measurements of the brain's immediate response to stimuli. Participants wear an EEG cap while they are exposed to a variety of stimuli, such as images on a computer screen. These stimuli could be sounds, visual cues, or other brief experiences that the participant can detect. The EEG cap holds electrodes that record the voltages at several locations on the participant's scalp, hundreds of times per second. EEG hardware varies in terms of the maximum number of samples per second and the number of scalp electrodes.

Changes in the voltage at the scalp are a result of the aggregation of many neurons firing [6]. For this reason, we expect the true underlying waveform to be smooth. While there is considerable noise when collecting any given timepoint of a single trial, we do not expect dramatic voltage changes from one timepoint to the next. Instead, we expect to see the voltage gradually rise and fall in the shape of anticipated ERP components with single peaks. These components are brief, systematic fluctuations in the measured electrical potential. Their well-studied hill shapes have led to researchers' interest in the maximum. However, some researchers choose to test the

single value of the maximum amplitude directly, while others use the average of the values in a prespecified window to represent the entire component. Both approaches are used to test for differences in how brains respond to stimuli.

The following are the typical steps, or pipeline, for analyzing ERPs, in order. They have been adapted from several manuals, including ones by Cohen [10], Keil et al. [37], and Luck [42]. These steps will be referenced by number throughout later sections of this paper.

1. Collect data in a continuous stream using EEG, making note of the time of stimulus onset for each trial, the trial condition, and any other important event codes (such as correct or incorrect response and response onset).

2. Clean data:

    a) Filter data to remove long-term trends or drift [for more information on filters, see 11].

    b) Remove or correct artifacts (such as eye blinks, sneezes, coughs, etc) using, for example, regression or ICA [see 37, p. 6, for a full list and references].

    c) Optionally, re-reference the data. Standard references are a mastoid or the average of all sensors. The choice of reference electrode can be impactful [16], but can be leveraged to explore spatial relationships [35].

    d) Epoch the continuous data to create single-trial EEG segments (e.g., from -200 to +800 milliseconds).

    e) Baseline each trial so that voltages are relative to voltages prior to the key stimulus or response onset. As a result, each trial begins at zero microvolts.

3. Average the single-trial EEG epochs (by taking the average at each time point) to create single-subject averaged ERP waveforms for each condition of interest.

Averaging is done here to gain a higher signal-to-noise ratio, by averaging out the variability of individual trials [10].

   a) Some researchers take difference waves to compare two conditions [see, for example, 38].

   b) Low-pass filters may be applied here. Low-pass filters attenuate high-frequency signals in order to increase the signal-to-noise ratio, in much the same way that averaging over trials smooths out the waveform. Thus, averaging over trials can be considered a specific kind of low-pass filter.

4. Using a prespecified window to isolate the component of interest, calculate the summary measure (either maximum or average) for each condition-level average waveform for each individual. This step reduces the average waveform for each subject-condition pair to a single value.

5. Perform a within-subjects ANOVA or paired t-test for group-level analysis.

   a) Corrections such as Greenhouse-Geisser are commonly used for omnibus tests.

While these steps outline the standard for ERP research, there are many decisions left to the researcher within this process. In this paper, we focus primarily on Step 4, the choice between using the maximum or the average as a summary statistic of the average waveform that results from Step 3. In doing so, we also address issues involving the averaging over trials that takes place in Step 3. This exploration will naturally involve other aspects of the study design and research plan, when these elements interact with the choice between maximum and average.

2.3 Comparison of Maximum and Average: An Example

To illustrate the potential differences in analyses based on either the maximum or the average, we begin with an example of a simulated study of a single component. Our goal in this section is to demonstrate the impact of decisions such as use of the maximum or average, and to highlight related issues that might emerge within the traditional analytic pipeline of ERP data. We mimic a study in which 20 participants experience both common (30 trials) and rare (10 trials) stimuli and are compared across these two conditions on the P300 component following the pipeline description in the previous section. This study design is similar to, for example, the oddball paradigm of the novelty P300 study by Friedman, Cycowicz, and Gaeta [26]. This simulation produces differences on the P300 component only and ignores the rest of the waveform, so that we can limit our investigation to only this component without contamination from other, potentially overlapping, components. For analysis, we make use of the standard ERP pipeline detailed in Section 2.2.

A common step for presentation is to plot the grand-averaged waveforms, so that the two conditions can be compared qualitatively. Figure 2.2 shows how responses to the rare and common stimuli compare to one another. To create this figure we followed standard procedure and first averaged all trials for each condition within person, and then averaged over people. In this case, order of operations does not matter because we have no missing trials.

We can see in Figure 2.2 that the two conditions appear to have some differences. The common condition has a somewhat lower voltage throughout the area of interest (near 300 milliseconds). It is difficult to tell if this difference is significant or not because the plot has no representation of the underlying variability at each time point. Instead, this plot only shows the stability of the process over time. We may also notice a slight delay in the peak of the common condition relative to the rare condition, and a potentially overlapping component in the rare condition that may

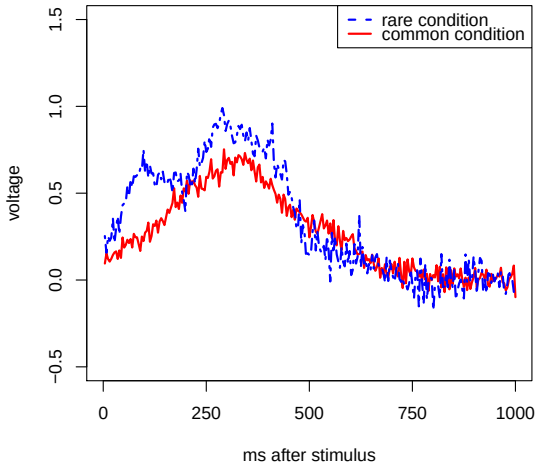


Figure 2.2: Grand average waveforms comparing simulated P300s

be contaminating the signal.

When we focus on only observations in our window of interest (250-350 milliseconds post-stimulus), we can come close to picking the maximum for each condition, and might be able to make a reasonable guess at the average. Table 2.1 shows the values for the maximum and average voltages in each condition, based on the grand average waveforms.

	maximum	average
common condition	.7527	.6593
rare condition	.9906	.8626

Table 2.1: Comparison of maximum and average from grand-averaged waveforms

This table highlights two important details. First, the maximum is greater than the average in each condition. This will always be true, as it is always true that the maximum is as large or larger than the average in any set of values. (The only way for the average to be as large as the maximum is for all values to be the same. Any values that are not as large as the maximum reduce the average while leaving the maximum unchanged.) The table also confirms what was already visible in Figure 2.2: in this case, the rare condition has larger values than the common condition in

both metrics. This table, like the figure, does not give us any idea of the variability in the underlying data because it is based on only the two grand-averaged waveforms.

In order to test if the difference is statistically significant, we can perform a paired t-test using the summary measures from the waveforms that result when we average all trials for each condition within subject. We have 2 summary measures for each person, and thus can measure intersubject variability of the differences and perform a significance test. We lose the intrasubject variability information by averaging over trials in Step 3. Table 2.2 shows the results of the 2 paired t-tests for the maximum and average.

Test	Results
Paired t-test for maxima:	$t(19) = 4.16$, $d = 1.31$, $p = .0005^{***}$
Paired t-test for averages:	$t(19) = 1.56$, $d = 0.49$, $p = .136$

Table 2.2: Comparison of t-tests for maxima and averages across conditions

We can see that the maxima and averages yield different test results for this particular simulation study. The data were generated to have greater amplitude in the rare condition than the common one, which only the test using the maxima finds. The amplitude difference in the simulation is designed to be approximately 1 microvolt, with a 50-millisecond delay in the common case, so the difference was overestimated when using maxima and underestimated when using averages. Now that we know the design and result, we can examine this particular study and highlight some of the contributing factors that might lead to this disparity in the statistical results.

First of all, not only will the maximum and average take different values, but when viewed as random quantities, they have different underlying distributions. Using assumptions that agree with the context of ERP research, we derive properties of the underlying distributions of these quantities and pairwise differences of these quantities in Section 2.4. Luck [42] argues that the order of operations will not impact final results, so we also discuss whether the underlying distributions are invariant to

reordering of Steps 3-4.

In the grand-averaged waveforms in Figure 2.2, there are a few small spikes that create local maxima. These are much more frequent and dramatic in the within-person waveforms—Figure 2.3 shows a comparison of the averaged waveforms for the two conditions for just one participant. The spikes, artifacts from the variability of each trial, allow for occasional large values for the maximum. Assuming that the noise across observed timepoints is symmetrically distributed, we can expect the observed maximum to be positively biased with a bias proportional to the variance within that single timepoint. The average, however, should not be biased. This can have a major impact on statistical tests. To get a large difference within person for the maxima, only one large spike is needed, whereas a large difference in the average requires a more systematic difference in the two conditions throughout the window of interest.

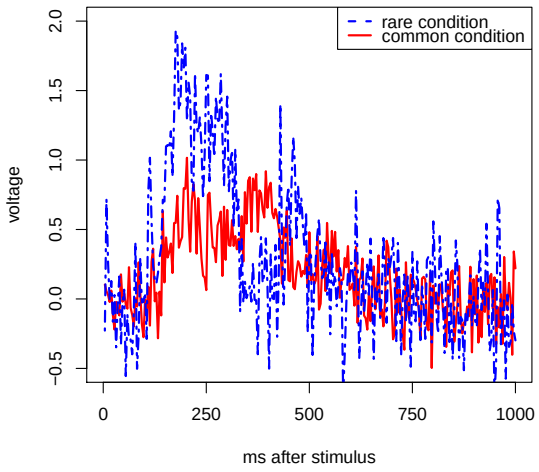


Figure 2.3: Waveforms comparing two conditions within a single subject

Because outliers can be particularly influential to the maximum, perhaps a larger sample may have helped. ERP studies have a hierarchy of trials, subjects, and groups. This hierarchy could be included in the model (e.g., by using random effects), but this is rarely done. A benefit of such an analysis is that one can model the multiple sources

of variability (e.g., test for and model heterogeneity in intrasubject variability). This raises the question of the optimal number of observations and how they should be allocated at every level of the study.

We can repeat the simulated study using a range of subject sample sizes. Table 2.3 shows that a larger sample size helps the average to identify the difference, but the maximum has consistently higher test statistic values (i.e., smaller p-values) for identifying differences. This may be counterintuitive given that the average uses more information than the maximum. We will explore both Type I and Type II error rates in more detail in Section 2.5.

N	Test	Results
10	Paired t-test for maxima:	t(9) = 2.72, d = 1.22, p = .023*
	Paired t-test for averages:	t(9) = 1.08, d = 0.48, p = .307
20	Paired t-test for maxima:	t(19) = 4.16, d = 1.31, p = .0005***
	Paired t-test for averages:	t(19) = 1.56, d = 0.49, p = .136
30	Paired t-test for maxima:	t(29) = 5.30, d = 1.37, p = .00001***
	Paired t-test for averages:	t(29) = 3.40, d = 0.88, p = .002**
50	Paired t-test for maxima:	t(49) = 5.91, d = 1.18, p = .0000003***
	Paired t-test for averages:	t(49) = 2.40, d = 0.48, p = .020*

Table 2.3: Comparison of paired t-tests as subject sample size varies

There are other contributing factors that may interact with the choice of summary metric (either maximum or average) to impact statistical tests. Through a series of simulations in Section 2.5, we explore several of these elements, beginning with a *confirmation of distributional results* (see Section 2.5.1).

Another level at which we may vary the sample size is to create *unbalanced designs via trial counts* (see Section 2.5.2). In the current example, each person has 10 trials in the rare condition and 30 trials in the common condition. Because we take the average of the trials to remove noise, the unequal trial count may be leading to unequal variances that impact test results. We simulate studies in which the number of trials varies to explore this hypothesis.

While the single observation nearest the time when the true maximum occurs may

have negative noise, the dense sampling that is standard for ERP research will lead to several observations being collected near the true maximum, and we can expect one or more of these to have positive noise. The value near the maximum with positive noise is the one that will be identified as the sample maximum. The average value, however, will not be subject to this bias. We will also use simulations to see if varying the *sampling rates* (see Section 2.5.3) can help or hinder the standard testing framework in identifying which spikes are true signals and which are due to noise.

Another possible factor in the difference in results is the small latency differential that appears to exist between the two conditions. Indeed, the simulation included a slight delay in the common condition. While it does not look like much of a difference at this scale, the discrepancy can be more dramatic at the individual level. A latency difference might mean that the window location was not optimized for both metrics. A small miscalculation on the window location will not make a difference in the maximum because as long as the value is in the window, it will be used. The average, however, will change with small adjustments of the window. If the component is symmetric and hill-shaped, we would want the window to be centered on the peak to get the largest value for the average. Also, a small window will yield the largest value for the average. Luck [42] makes different recommendations for window size based on summary metric—smaller windows when using the maximum than when using the average, and a window size of at least 40 milliseconds for the average. We use simulations to explore error rates in cases where we have *latency differences across window locations and widths* (see Section 2.5.4).

Before we get to the simulations, we review some basic and relevant distributional theory. These theoretical results will inform the simulation in the subsequent section.

2.4 Distributional Theory of Summary Metrics

When we compare tests using local maxima versus averages, our primary concern rests with the underlying distributions of these two quantities to ensure that the distributional assumptions of the test match the properties of the quantities of interest. In most ERP studies, ANOVAs are used to compare groups of interest. To simplify the comparison, we will focus on the simple two-condition version of repeated-measures ANOVA, the paired t-test. One of the assumptions of this test is that the differences are normally distributed. It is not intuitively obvious if this assumption is met when testing maxima. We also must consider how the variance differences outlined in Section 2.3 come about and how they will impact statistical tests. In this section, we use statistical theory to explore the asymptotic distributions and convergence rates of summary metrics as they relate to ERP testing.

One substantial assumption is necessary to simplify the exploration of underlying distributions. We assume that all observations in a given time window are independent and identically distributed (IID). While observations of an EEG recording are clearly not temporally independent, the maximum and average do not make use of the temporal ordering of the data. For this reason, we proceed under the assumption that independence is plausible. We assume that each time point has noise that is produced in a consistent way, as a combination of human physiology and the EEG equipment. Thus the observations are identically distributed. Because we first average at each time point over all trials for a condition (we can also assume that the sampling rate is exact to simplify this step), it does not matter how the individual time points are distributed. Under the central limit theorem, each time point is normally distributed with variance proportional to the amount of noise that is always present in ERP data due to the way it is collected.

For one window on a single condition, we have a sample size determined by the window size and the sampling rate, as described in Equation 2.1. For example, if our

window is two tenths of a second and we use a 256 Hz sampling rate, $n = 61$. If our window is smaller, say one tenth of a second, and the sampling rate lower, such as 128 Hz, $n = 12$. As we will see in Section 2.5, these seemingly small choices can lead to substantial impacts later on, when we are depending on asymptotic results.

$$n = \lfloor (\text{window duration in seconds}) * (\text{sampling rate in Hz}) \rfloor \tag{2.1}$$

The distributional theory for the average is simple. The central limit theorem tells us that the sum, and thus the average, of n IID observations tends towards a normal distribution as n tends to infinity. The rate of convergence is generally stated to be $\frac{1}{\sqrt{n}}$, and if the original distribution is already close to normal, only a small n is needed to make tests with a normal distributional assumption valid. These rules of thumb can be explored more thoroughly by using the Berry-Esseen Theorem.

The Berry-Esseen theorem helps us to estimate the sample size n needed for reasonable convergence. The distance between F, the cumulative density function (CDF) of the IID samples, and a normal distribution is:

$$D \leq \frac{C\rho}{\sigma^3\sqrt{n}}, \tag{2.2}$$

where ρ is $E[|X_1|^3]$, the third absolute moment, σ is the standard devation of F, and C is a constant less than .56 [57] and greater than .41 [23].

A limitation of using this theorem to quantify our fit to distributional assumptions is that the Berry-Esseen Theorem only gives an upper bound on the distance. That is, it gives the worst-case scenario of the distance from normality. It is entirely possible that a distribution yielding a larger right hand side of Equation 2.2 can be closer to normality than one with a smaller right hand side.

The two main elements that affect this upper bound are the sample size, n, and the ratio ρ/σ^3. The $\sqrt{n}$ in the denominator of Equation 2.2 is where our usual rule

of thumb comes from. However, this equation shows that ρ and σ also affect the convergence rate. The third absolute moment, ρ, can be affected by many aspects of F, but especially by the symmetry. When F is not symmetric, ρ becomes large, but σ also typically grows as F becomes less symmetric. Thus it is challenging to study ρ and σ separately.

Regardless of the rate of convergence, our asymptotic theory shows that when we work with averages within a window we ultimately will have a normal distribution that meets the distributional assumptions of t-tests and ANOVAs. Equation 2.3 describes this asymptotic result from the central limit theorem.

$$\sum_{i=1}^{n} \frac{X_i}{n} \sim N(\sum_{i=1}^{n} \frac{x_i}{n}, \frac{s^2}{n}) \tag{2.3}$$

The distributional theory for the maximum is not as simple. If the true underlying component has a well-defined peak, then the maximum happens at (or near) that point. Because each timepoint is first averaged across trials within each person, this single point will have a normal distribution and thus the maximum (and the difference of maxima) may be normally distributed. However, if the component achieves a plateau rather than a peak, the timepoints within the span of similar observations may be nearly identically distributed and may also be treated as independent. In this case, just as the central limit theorem is used to describe the asymptotic distribution of sample means, the Fisher-Tippett-Gnedenko Theorem (also known as the extreme value theorem) can be used to tell us about the asymptotic distribution of the sample maximum. This seemingly trivial property of the data-generating function shifts the distribution of summary statistic—a concern that we return to in Section 2.6.

The Fisher-Tippett-Gnedenko Theorem states that, if the distribution of the maximum converges, it must converge to one of three distributions: Gumbel, Frechet, or Weibull [25, 29]. For convenience, these three distributions have been generalized as special cases of the Generalized Extreme Value (GEV) distribution. The theorem

does not guarantee convergence or give criteria for convergence, but we will assume here for simplicity that the distribution does converge. The simulations in the next section will help to assess the validity of this assumption, and possible rules of thumb for necessary sample sizes.

More formally, the Fisher-Tippett-Gnedenko Theorem states:

Theorem 2.1. *Let $X_1, X_2, ..., X_n$ be a sequence of IID random variables. $M_n = \max\{X_1, X_2, ..., X_n\} = X_{[n]}$. If there exist a pair (a_n, b_n) such that each $a_n > 0$ and $\lim_{n\to\infty} P(\frac{M_n - b_n}{a_n} \leq x) = F(x)$, then the CDF $F(x)$ is Gumbel, Frechet, or Weibull.*

Additionally, Gnedenko [29] outlined the domains of attraction for each of the three possible GEV distributions. For example, the maximum of a series of IID standard normal distributions follows a Gumbel distribution. The details of this theory has been described in various texts and articles, such as one by de Haan [15]. Because we first take the average over many trials, each time point should asymptotically have a normal distribution, possibly with different means. Thus it seems that a Gumbel distribution would be a reasonable choice for modeling the maximum. Simulations in the next section explore when a Gumbel distribution will result when taking the maximum.

The Gumbel (μ, β) probability density function (PDF) is:

$$\frac{1}{\beta} e^{-(z+e^{-z})} \qquad \text{where } z = \frac{x-\mu}{\beta}.$$

If we want to compare the maximum M_{n1} from one condition with the maximum M_{n2} of another condition, and we assume that the scale parameter β is the same for each underlying distribution, we can make use of the fact that the difference of two Gumbel-distributed random variables with the same variance follows a logistic distribution. When we test that the location parameters μ_1 and μ_2 for the two trials are equal, the null distribution of $M_{n1} - M_{n2}$ follows logistic$(0, \beta)$. The proof, which

makes use of u-substitution, follows:

Proof. Let $Z = X - Y$, where $X \sim$ Gumbel (a, β) and $Y \sim$ Gumbel (c, β). F denotes a CDF, while f denotes a PDF.

$$
\begin{aligned}
F_Z(z) &= P(Z \leq z) \\
&= P(X - Y \leq z) \\
&= P(X \leq Y + z) \\
&= \int_{-\infty}^{\infty} F_X(z - y) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} e^{-e^{-\frac{y+z-a}{\beta}}} \frac{1}{\beta} e^{-(\frac{y-c}{\beta} + e^{-\frac{y-c}{\beta}})} dy \\
&= \int_{-\infty}^{\infty} e^{-e^{-\frac{y+z-a}{\beta}}} \frac{1}{\beta} e^{-(\frac{y+z-a-z+a-c}{\beta} + e^{-\frac{y+z-a-z+a-c}{\beta}})} dy \\
&= \int_{-\infty}^{0} e^{u} e^{(-\frac{-z+a-c}{\beta} + u e^{-\frac{-z+a-c}{\beta}})} du \\
&= e^{-\frac{-z+a-c}{\beta}} \int_{-\infty}^{0} e^{u(1+e^{-\frac{-z+a-c}{\beta}})} du \\
&= \frac{e^{-\frac{-z+a-c}{\beta}}}{1 + e^{-\frac{-z+a-c}{\beta}}} e^{u(1+e^{-\frac{-z+a-c}{\beta}})} \Big|_{u=-\infty}^{0} \\
&= \frac{e^{-\frac{-z+a-c}{\beta}}}{1 + e^{-\frac{-z+a-c}{\beta}}} (1 - 0) \\
&= \frac{1}{1 + e^{\frac{-z+a-c}{\beta}}} \\
&= \frac{1}{1 + e^{-\frac{z-(a-c)}{\beta}}}
\end{aligned}
$$

This is the CDF for a logistic$(a - c, \beta)$ distribution. Under the null hypothesis, $a - c = 0$. □

The logistic(μ, s) PDF is:

$$\frac{e^{-z}}{s(1+e^{-z})^2}, \qquad \text{where } z = \frac{x-\mu}{s}. \tag{2.4}$$

s can also be written in terms of the standard deviation, σ:

$$s = \frac{\sqrt{3}}{\pi}\sigma.$$

The logistic and normal distributions are very similar [2]. The logistic distribution is slightly more peaked and does have slightly wider tails than the normal distribution [9]. However, the logistic distribution has been used in place of the normal distribution due to its relative simplicity [34, ch. 22]. This history suggests that in an applied setting, particularly when we have already made so many assumptions and are already dealing with an approximation to a distribution (due to asymptotic theory), it should not be problematic to use a normal distribution to test pairwise differences between maxima of components. Indeed, simulations in Section 2.5 also show that this difference looks logistically or normally distributed.

While the Berry-Esseen Theorem gives convergence rates to normality, there is not equivalent theory to suggest a convergence rate for the convergence to Gumbel (and thus, of the differences to logistic). Our simulations will attempt to show the effects of various study design decisions on this convergence.

The order of operations is, at least from a theoretical perspective, quite important. If our methodology involved taking the maximum of each trial and then averaging over all trials in each condition for each person, then we would be working with averages, and thus the central limit theorem, the Berry-Esseen Theorem, and normality. This paper would become primarily a discussion of convergence rates, instead. When working with the average as the summary statistic, we take means twice: in Step 3 (to average the waveform) and Step 4 (to summarize the values in the window of interest).

The average would be unaffected by changing the order of these steps, assuming equal sample sizes or proper weighting of observations in an unbalanced design. Our first step would still result in a variance of $\frac{\sigma^2}{n}$ when describing the distribution of the summary statistic. However, the maximum would have an asymptotic variance of $\frac{\sigma^2\pi^2}{12\log(n)}$ [60]. Because the denominator grows with n at a faster rate for the average than the maximum, convergence of the variance will be faster for the average. The next step will involve averaging over the summary statistics, which will asymptotically yield a normal distribution. Looking at Equation 2.2, unless ρ is much larger for the maximum case than the average case, convergence to the normal distribution will be faster for the average than maximum.

The following section will describe a simulation framework that will allow us to both ensure that distributional assumptions are met and explore related issues in the design and analysis of ERP studies, as raised in Section 2.3.

2.5 Simulating Data from an ERP Component

Because the underlying theory for the distribution of the maximum is complex and differs substantially from that of the average, simulations can provide an additional venue for exploring the practical impacts of various study design and analysis choices. The simulation described here allows us to isolate dynamic aspects of an ERP study and explore how the elements described at the end of Section 2.3 might impact statistical analyses in the long run.

The simulation focuses on only one ERP component. To achieve a hill-shaped component, we use a normal kernel. As noted in Section 2.4, the shape assumptions made here impact the underlying distributions of our summary statistics. The choice of normal kernel has been used by Helwig [32] as a way to accurately recreate visual-stimuli ERP waveforms based on data from several studies. Helwig's `eegsim` function in the `eegkit` package in `R` uses a prespecified voltage weight for each channel and

multiplies the functional form of a normal kernel by this weight to simulate ERP components. For example, a P300 at the P5 electrode at T milliseconds after stimulus onset is simulated as

$$4.79e^{-300(T-.35)^2}.$$

We first generalize this process in Equation 2.5, then specify values for our simulation. The following is, step-by-step, how a simulated study is generated. The starting point for all of the waveforms is:

$$v_i = He^{-W(t_i-L)^2} \tag{2.5}$$

where H determines the height of the component, W determines the width, t_i is the time (in seconds) post-stimulus, and L is the latency of the peak of the component. This shape returns to 0 on either side of the hill-shaped component, which reflects baselined data. Thus we can assume that the baselining step has already been performed when we use this simulation. For this paper, we model our true underlying component after the P300 and look at times between stimulus onset and one second later. We set the true latency, or time of peak, for this component at 300 milliseconds, or .300 seconds after stimulus onset (time $t_i = 0$). Thus, we can plug in .300 for L:

$$v_i = He^{-W(t_i-.300)^2}.$$

Components are generally thought to have a consistent width for this hill shape, so the width is also fixed in all simulations to a value that was selected based on experience and discussions with ERP researchers:

$$v_i = He^{-200(t_i-.300)^2}.$$

The height H is set at 4 to reflect a peak amplitude of 4 microvolts, again based on experience and comparison to similar studies. This value simply sets the scale of the

y-axis, and is thus fairly arbitrary.

$$v_i = 4e^{-200(t_i - .300)^2}$$

This simulation is set up to reflect a simple study in which each participant experiences two different conditions, such as a go/no-go task. We assume that each condition produces specific modifications to this baseline waveform. We model this by adding small perturbations to both the height and latency of the component. Here, condition 1 yields a smaller, later response (like the common stimuli in Section 2.3) at 350 milliseconds and condition 2 yields a larger response centered at 300 milliseconds (like the rare stimuli in Section 2.3):

$$v_{i1} = (3)e^{-200(t_i - .350)^2}$$
$$v_{i2} = (5)e^{-200(t_i - .300)^2}.$$

Each person is slightly different, so we allow for individual differences for each subject j in terms of both the height of the waveform and the latency:

$$v_{i1j} = (3 + h_j)e^{-200(t_i - (.350 + l_j))^2}$$
$$v_{i2j} = (5 + h_j)e^{-200(t_i - (.300 + l_j))^2}$$

where $h_j \sim N(0, 1)$ and $l_j \sim N(0, .100)$. We assume that these individual differences are the same across conditions within person, so there should not be a person-by-condition interaction. The choice of standard deviation for l_j is based on an assumption that the majority (68%) of individuals will have a true peak latency for the P300 component between 200 and 400 milliseconds post-stimulus. While latency is often not published for studies focused on analyzing amplitudes, Michalewski, Prasher, and Starr [45] estimated a standard deviation of 20-25 milliseconds for the P300

component. There are many possible explanations for this—it may be that the true latency variability across individuals is this small, or this value may be confounded by partially overlapping components, the filtering process, or other aspects of the data collection and analysis procedure. In the following sections, we explore the impact that this latency variability across subjects has on our simulation results.

Trial variability for each trial k within each condition is modeled similarly. While the trial variability could be specific to the individual, we assume here that it is not. Thus, the component is generated as follows for each trial in the study:

$$v_{i1jk} = (3 + h_j + h_k)e^{-200(t_i-(.350+l_j+l_k))^2}$$
$$v_{i2jk} = (5 + h_j + h_k)e^{-200(t_i-(.300+l_j+l_k))^2}$$

where $h_k \sim N(0, 1)$ and $l_k \sim N(0, .100)$. Again, we try both .100 and .020 as possible values for the standard deviation of l_k, matching the choice used in l_j. However, Step 3 (averaging over trials within person) causes this choice to have little impact.

The last area in which noise is introduced is at each time point of each trial. We assume that most of the noise at a specific timepoint is a product of the equipment and overall human physiology rather than the subject or condition, so this noise $z_i \sim N(0, 1)$ and is thus independent of anything being manipulated in the study. We arrive at the final model used to simulate voltages at time i of trial k for subject j in each of two conditions:

$$v_{i1jk} = (3 + h_j + h_k)e^{-200(t_i-(.350+l_j+l_k))^2} + z_i \tag{2.6}$$
$$v_{i2jk} = (5 + h_j + h_k)e^{-200(t_i-(.300+l_j+l_k))^2} + z_i. \tag{2.7}$$

This simulation design allows for variation at each of the levels of the naturally-occurring hierarchy, but has been designed to exclude interactions across these levels. For example, the trial variability is not related to the condition or the person. By

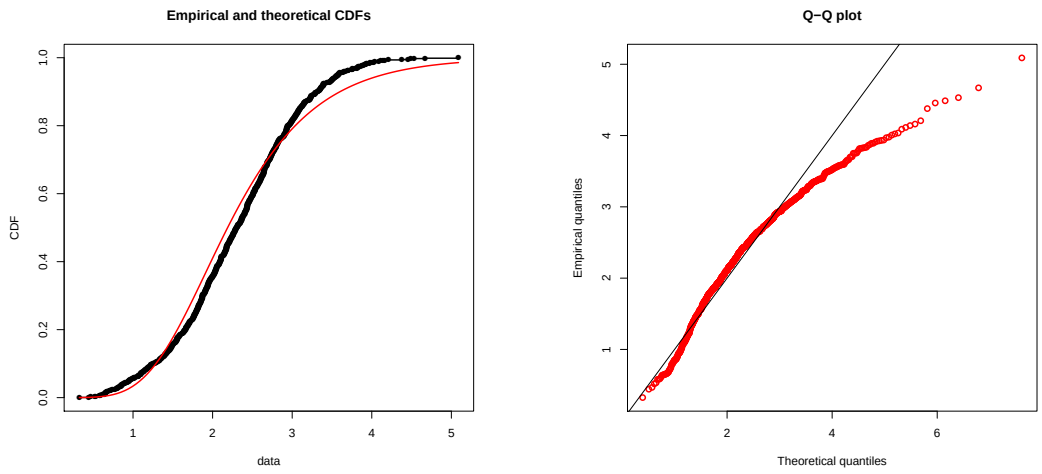
designing the simulation in this way, we can manipulate individual variables while maintaining an ecologically valid model.

There are many variables that remain unstated at this point. These variables reflect aspects of the study design, whereas the definitions above outline the model for the components themselves. For example, the sampling rate (measured in Hertz, or observations per second) will determine the timepoints t_i at which data is available. We assume that the timepoints are equally spaced, start at stimulus onset, and end at one second; that is, if we sampled at 5 Hertz, we would have samples at $t_i = \{0, .25, .5, .75, 1\}$. We can also vary the number of trials per condition and the number of subjects in the study. The remaining element that can be varied is the size and location of the window used to calculate the summary statistic (either maximum or average). We assume for simplicity that no trials are dropped from the study, although this is not normally the case in ERP studies. Simulations make use of 1000 repetitions, unless otherwise stated.

The results of the simulations comparing maximum and average summary measures in ERP data are organized into four subsections. The first subsection checks distributional forms, the second subsection examines the role of the number of trials on Type I and Type II error rates, the third subsection focuses on the role of the sampling rate on Type I and Type II error rates, and the fourth subsection examines latency differences, window locations, and different window sizes on Type I and Type II error rates.

2.5.1 Confirmation of Distributional Results

As discussed in Section 2.4, the shape of the underlying component will affect whether the maximum follows a normal or Gumbel distribution. In order to explore the distribution of the maximum, and confirm that assumptions are met for testing the difference of maxima using a t-test or ANOVA framework, we followed the procedure



(a) CDF (b) Q-Q Plot

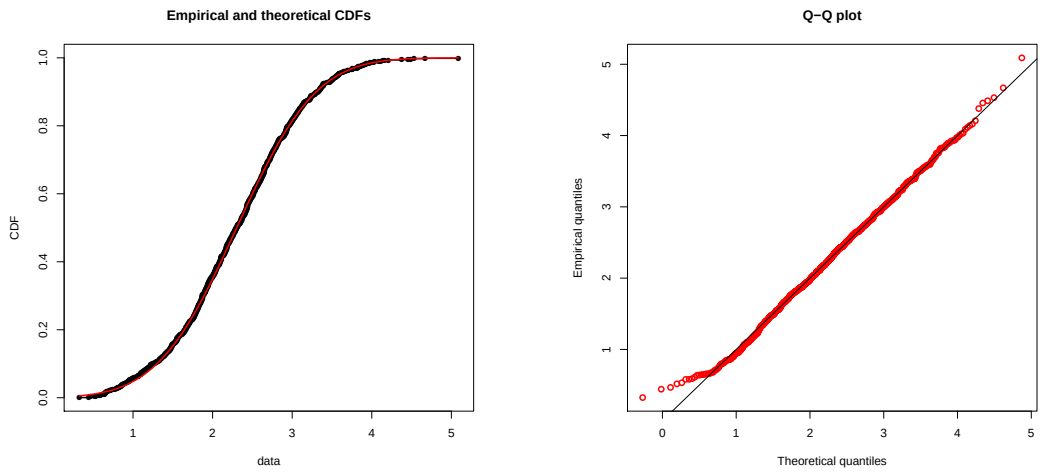
Figure 2.4: Distribution of maximum compared to Gumbel distribution

outlined in Steps 1-4 of standard ERP analysis using the approach above. To start, we generated 20 trials for condition 2 for a single person, then found the maximum between 250 and 350 milliseconds. We repeated this process (keeping the same person values h_j and l_j in condition 2) 1000 times.

We first fit a Gumbel distribution to the data using maximum likelihood. As shown in Figure 2.4, a Gumbel distribution provides a reasonable fit to the data. The data appears somewhat skewed to the right in Figure 2.4a, but Figure 2.4b highlights that the tail is not as heavy as it should be with a Gumbel distribution.

Figure 2.5 shows a reasonable fit for a normal distribution to the data as well. One possible reason for this is that the simulation is designed with a single point where the wave peaks (rather than a flat plateau over many milliseconds), and the maximum occurs at that time. Because the point is generated with normally distributed noise (from both amplitude and latency variation), the distribution of the maximum is close to normal in this case. These distributional results also suggest that our IID assumption in the previous section may have been too strong.

The underlying ERP analytic plan makes asymptotic statements, including one that the difference between the maxima of the averages of many trials will be approx-



(a) CDF (b) Q-Q Plot

Figure 2.5: Distribution of maximum compared to normal distribution

imately normal. Following the simulation above, we computed the difference between the maxima across the two conditions as one would do in a typical ERP study. We plotted the differences between 10,000 pairs. Figure 2.6 shows that the difference of maxima does appear to be normally distributed. Thus the distributional assumption for pairwise testing and ANOVAs is met, regardless of the distribution of the maximum for a single condition.

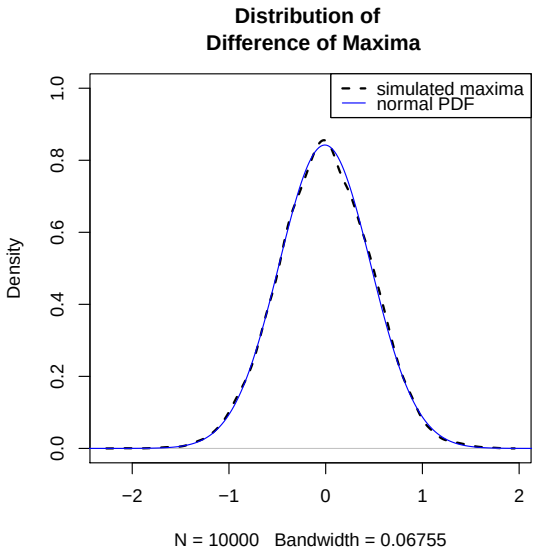


Figure 2.6: The distribution of the differences in maxima is approximately normal

2.5.2 Unbalanced Designs via Trial Counts

The example study in Section 2.3 involved unequal numbers of trials in the two conditions. Studies with unbalanced designs, such as those from oddball paradigm like that of Friedman et al. [26], are common. In this subsection, we use a series of simulations to explore the impact of the number of trials per condition on results.

First, we simulate a study similar to that of Section 2.3, where there is a true difference in amplitude, keeping the number of condition 1 trials (generated using Equation 2.6) fixed while the number of trials in condition 2 (generated using Equation 2.7) varies from 5 to 80. We repeat 1000 studies, each with 20 subjects, to investigate the false negative rate.

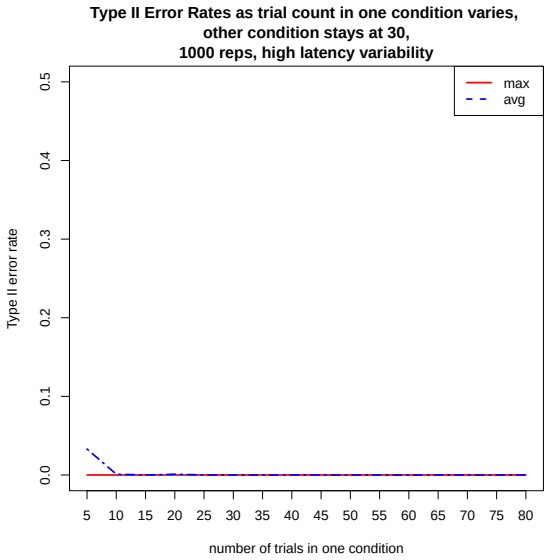


Figure 2.7: Type II errors in unbalanced designs

As we can see in Figure 2.7, both metrics yield low Type II error rates in studies with unbalanced trial counts across conditions. Like the example in Section 2.3, there were a couple errors where the average failed to yield significant differences when the trial count in one condition was only 5 per subject, but even these errors were rare.

Another way to approach this simulation is to consider the false positive rate. These Type I errors are assumed to be held at a researcher-specified constant. How-

ever, when this assumption does not hold, we find significant results in nonsignificant situations more often than we should. In this context, false positives can lead to researchers incorrectly claiming to have found differences between groups, and resources may be reallocated to studying an effect that does not exist.

To investigate the Type I error rate, we again simulated 1000 studies, but this time both conditions were generated from Equation 2.7 for condition 2. As with the false negative simulation, we keep the number of trials in one condition fixed at 30 while the number of trials in the other condition varies from 5 to 80. By doing so, we can compare the Type I error rate for paired t-tests using the maximum and the average.

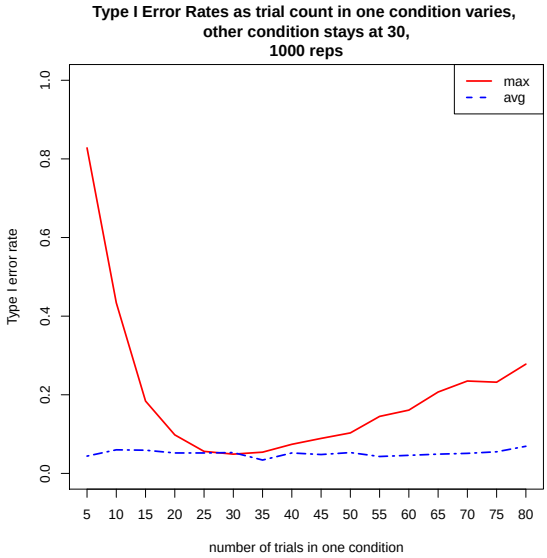


Figure 2.8: Type I errors in unbalanced designs

As Figure 2.8 shows, unbalanced designs can have dramatic impacts. The maximum finds many more false positives when the number of trials in the two conditions is not similar. To explain why this might occur, we can examine Figure 2.9. It shows the averaged waveform for one subject, using either 60 trials, a subset of 30 trials, or a subset of 5 trials. As the number of trials decreases, the signal-to-noise ratio decreases and the waveform has more peaks as a result of the smaller number of trials

being averaged at each timepoint. More peaks provide more opportunities to find a difference in the maxima, even when one does not exist.

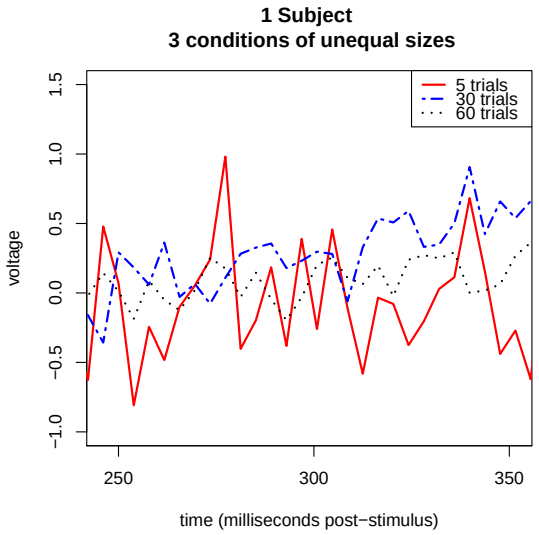
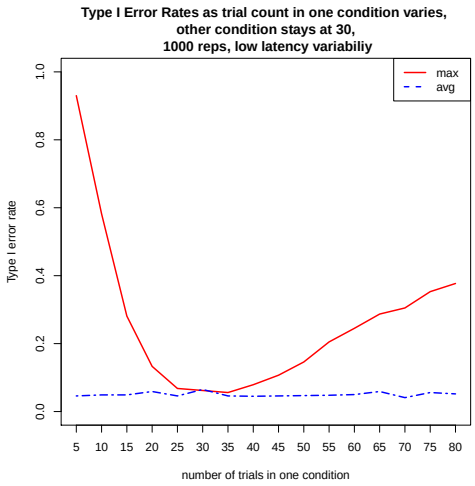


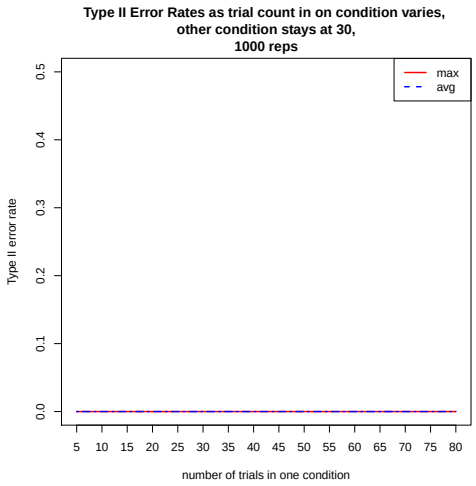
Figure 2.9: Prototypical components across differing trial counts

When we reduce the latency variability across individuals, a similar trend emerges in the Type I error rates. However, the error rates for the maximum remain slightly higher when the latency variability is reduced. This may suggest that higher variability trial-to-trial occasionally causes the small numbers of individual waveforms to combine to a less-pronounced peak for condition 2 [for a visual explanation, see Figure 4.2 in 42], leading to a lower amplitude and thus a non-significant difference. The Type II error rate is zero throughout. Results from the simulations with reduced latency variability are shown in Figure 2.10.

The simulations in this section suggest that the maximum may underperform in studies with unbalanced designs, particularly when no true difference exists. However, these simulations do not explain the results of our example study in Section 2.3 though the relatively high effective Type I error rate of the maximum (i.e., the maximum is too liberal) as demonstrated in this simulation may be one culprit.



(a) Type I Errors



(b) Type II Errors

Figure 2.10: Error rates in unbalanced designs with reduced subject variability

2.5.3 Sampling Rates

Another possible feature that may affect results is the sampling rate. Equation 2.1 describes how the number of observations being used to calculate the maximum or average relates to both the window size (which will be covered in Subsection 2.5.4) and sampling rate. As with the previous subsection, we simulated 1000 studies where both conditions were generated from Equation 2.7 for condition 2. We did this for various sampling rates.

We can see in Figure 2.11 that the Type I error rate for both methods appears relatively stable, with no differences between the two summary metrics, as sampling rate varies from relatively low (100Hz) to much higher than is standard for ERP studies (1000Hz). Figure 2.12 shows almost no Type II errors occurring, regardless of the use of summary metric. This suggests that our simulation was well-powered. Some authors propose that the choice of ERP sampling rate should be determined by the Nyquist rate [treated in more detail in 51]—the sampling rate should be at least double the rate of the phenomena of interest in order to reliably capture it. Because ERP components tend to last on the order of 100 milliseconds, a sampling rate of at

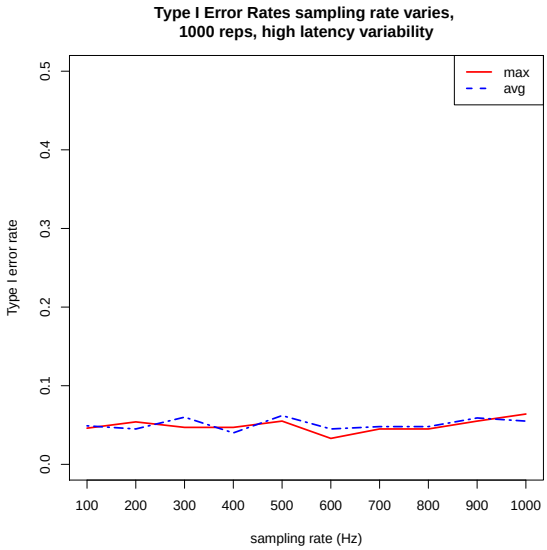


Figure 2.11: Type I error rates as sampling rate varies

least 100Hz should be more than adequate to capture it. If the component is more fleeting than the one in our simulation, lower sampling rates may fail to capture true differences and be underpowered.

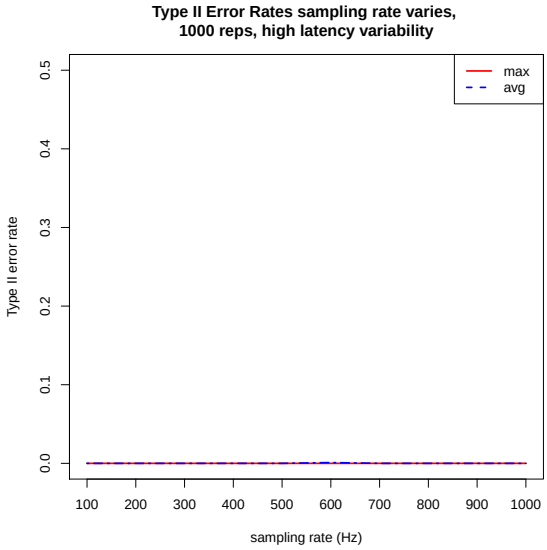
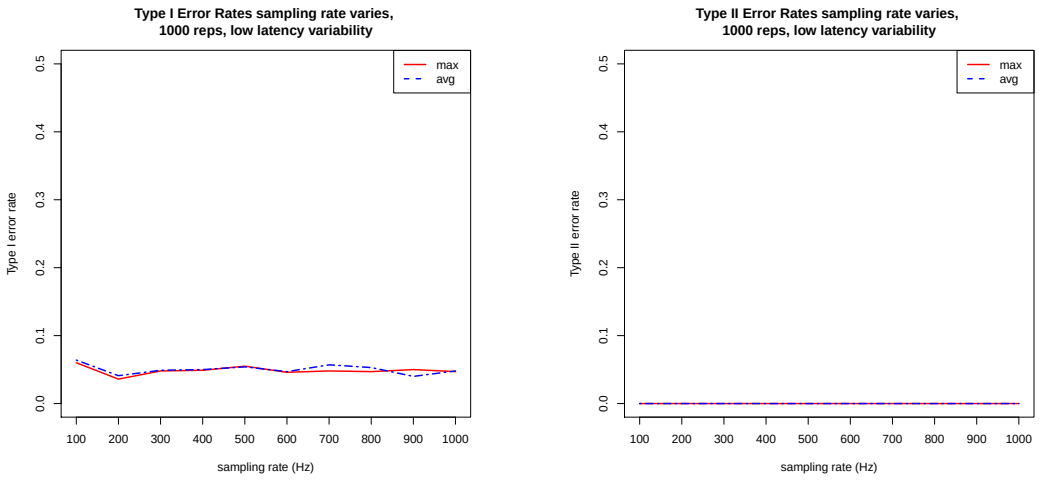


Figure 2.12: Type II error rates as sampling rate varies

Figure 2.13 shows that we find similar error rates with lower latency variability across individuals. Thus, it seems that sampling rate is not a major factor in the



(a) (b)

Figure 2.13: Error rates as sampling rate varies, low latency variability

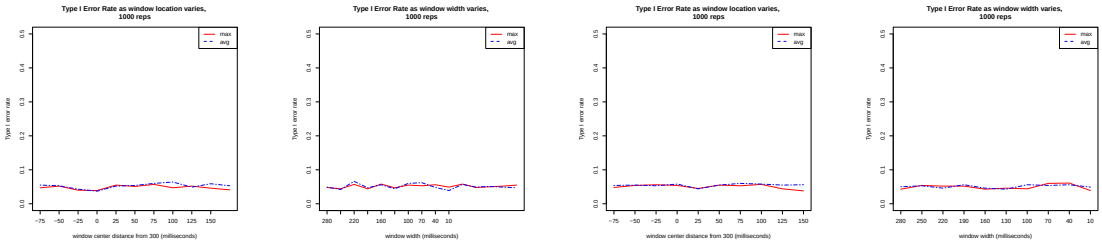
ERP experimental designs we have simulated in this paper.

2.5.4 Latency Differences Across Window Locations and Widths

As mentioned at the end of Section 2.3, the latency difference between conditions may impact results, even when it is not of direct interest. In this section, we work with two conditions with the same amplitude but different latencies. We assume that only the amplitude is of interest and explore how Type I error rates differ depending on the researcher's specifications for window width and location.

These issues have been briefly touched on in the literature, but not explored in depth. For example, Luck [42] recommends a larger window when working with the average than the maximum, with a width of around 40 for the average. Picton et al. [48] emphasize that increased variability in latencies will lead to smaller amplitude estimates.

We find that if there is no latency difference between the two conditions, the choice of window width and location does not matter. The Type I error rate remains near .05, even when the between-subjects variability is high. This is shown in Figure 2.14. However, if the two ERP waveforms differ in other ways, such as their trial-to-trial



(a) location varies, latency sd=.100 (b) width varies, latency sd=.100 (c) location varies, latency sd=.020 (d) width varies, latency sd=.020

Figure 2.14: Type I error rates as window size and location vary

amplitude variability or the latency variability of surrounding components, these findings may not hold.

We find that if there is a true latency difference between the two conditions, the window must be centered near the midpoint of the two condition peaks to minimize Type I error rate. Otherwise, the error rate becomes inflated at a rate relative to the between-subjects latency variability. Figure 2.15 shows that in the case where between-subjects latency variability is relatively low ($l_j \sim N(0, .020)$ and a true latency difference between conditions of 50 milliseconds), deviations from this ideal can lead to a rapid inflation of the Type I error rate.

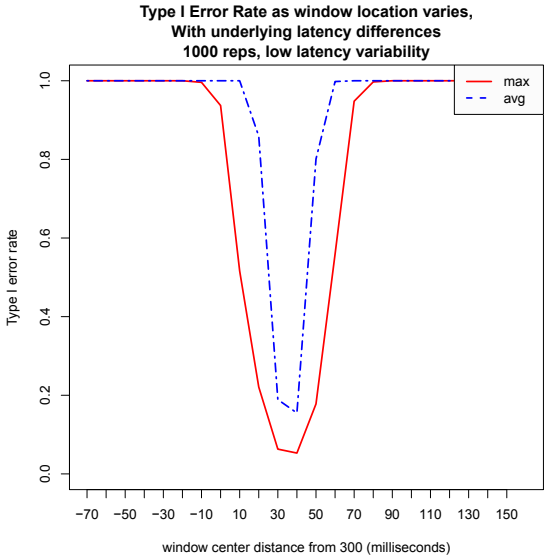


Figure 2.15: Type I error rates as window centering varies, for subjects with low latency variability

Figure 2.16 shows that when the between-subjects latency variability is higher ($l_j \sim N(0, .100)$ and a true latency difference between conditions of 50 milliseconds), deviations from the centering ideal again lead to inflation of Type I error rates, but at a slower rate than when the between-subjects latency variability is low.

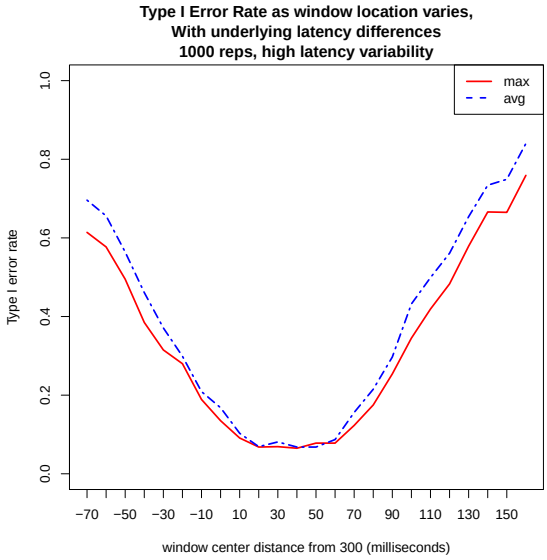
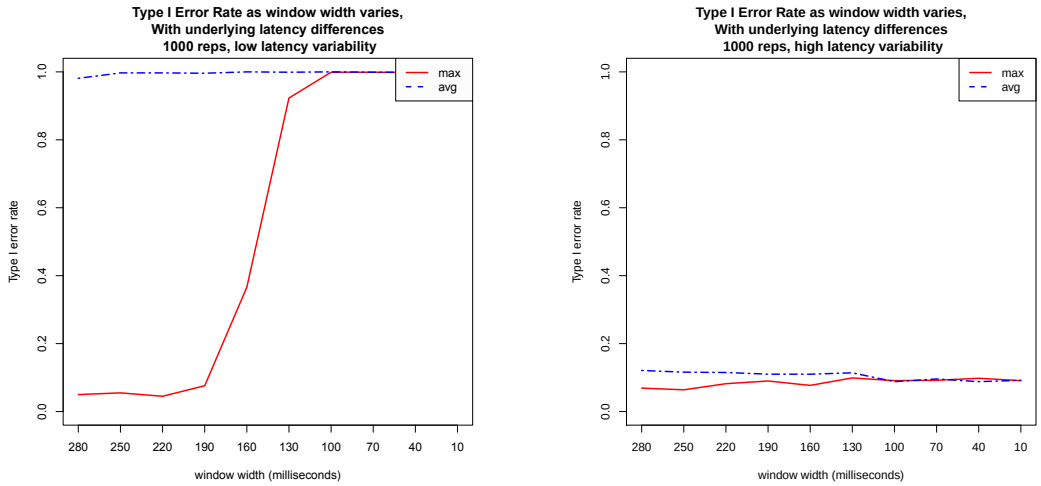


Figure 2.16: Type I error rates as window centering varies, for subjects with high latency variability

Regardless of the between-subjects variability, the maximum is more robust to errors in window location because any window that contains the true peaks, even if it does not contain the entirety of both components, will correctly test the amplitude difference. This also means that a wide enough window can overcome a non-ideal window location for the maximum, but not the average. Figure 2.17 shows how larger window widths interact with a non-ideal choice for the window center. In Figure 2.17a, the maximum maintains a Type I error rate near 5% for windows larger than 150 milliseconds because the window is large enough to cover both components consistently due to the small between-subjects latency variability. However, when the window becomes too small, the maximum begins missing the later condition's peak. The average has a very high error rate, as seen in Figure 2.15 at 300 milliseconds, and a larger window cannot help it recover. In Figure 2.17b, the error rate remains near

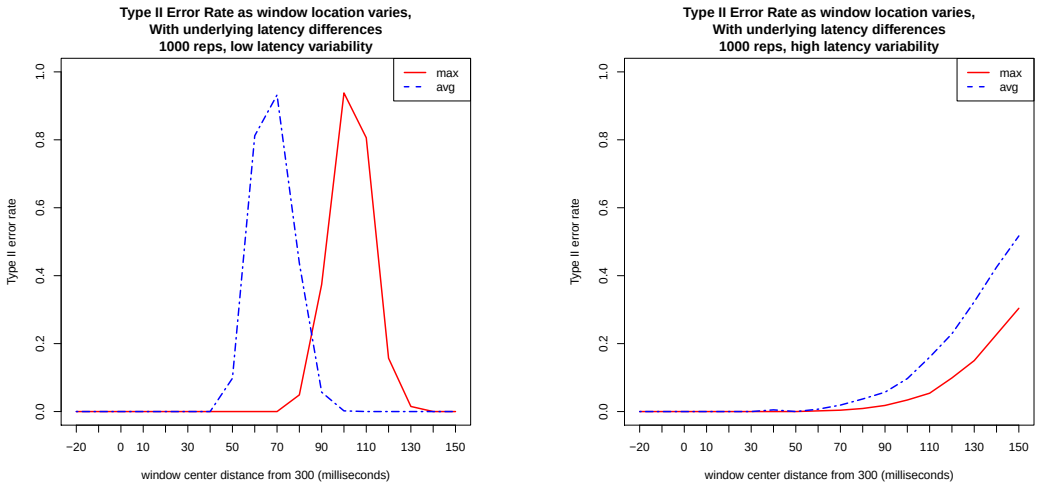


(a) sd= .020 (b) sd = .100

Figure 2.17: Type I Error rates as window width varies

10% for both measures because the high between-subjects latency variability leads to both components regularly falling within range of a window of any size centered at 300 milliseconds (note the 10% error rate at 300 milliseconds in Figure 2.16). In this case, the maximum again outperforms the average in large windows, but by a much smaller margin.

Inspecting the Type II error rates (Figure 2.18) reveals additional trends in errors that occur when the window is not centered at the ideal location. In this simulation, the peak of the condition with the larger amplitude occurred earlier. As the window center is moved to later times, the Type II error rate rises. As with the Type I error rate, the average is impacted first, then the maximum, and studies with smaller between-subjects variability see more rapid inflations of the error rates. The recovery of the Type II error rate at the latest timepoints of Figure 2.18a is due to the 2-sided testing procedure beginning to find that component 1 (the later component with lower amplitude) has greater amplitude than component 2. Because we know the true data-generating models in this simulation, we know that these are incorrect findings. However, in a real ERP study using ANOVA, this could easily be overlooked without careful inspection of the grand-averages waveforms and proper corrections in



(a) sd= .020 (b) sd = .100

Figure 2.18: Type II Error rates as window location varies

post-hoc testing. If it were the case that condition 2 had the larger amplitude, these plots would be mirrored across a vertical line at the time of the ideal window location.

Throughout these simulations, we explored many properties of these two summary metrics. Subsection 2.5.1 showed that the difference in maxima for a paired test follow a normal distribution, and thus meet the assumptions for standard ERP testing. Subsection 2.5.2 revealed that unbalanced trial counts lead to inflated Type I error rates for the maximum, but not the average. Subsection 2.5.3 found that sampling rate does not appear to impact results for either metric, as long as sampling rates are within the range of typical EEG equipment. Subsection 2.5.4 highlighted the importance of the window location when there is an unmodeled latency difference across conditions.

2.6 Discussion and Future Work

We explored the use of two common summary metrics in the context of ERP analysis. Despite concerns that knowledge of extreme value theory may raise, the use of local maxima in standard testing frameworks is not unreasonable if one's criterion

is matching distributional theory. We showed in Section 2.4 that the distributional assumptions for these tests will be met when using either the maximum or the average. However, our example in Section 2.3 showed that these two summary metrics can yield contradictory results.

To follow up on this observation, we used simulations to explore the differences in Type I and Type II error rates across several common scenarios. These simulations revealed a few valuable findings about the performance of these two summary metrics. First, the use of the maximum is more likely to lead to a false positive finding when comparing groups in unbalanced designs, such as oddball paradigms (as shown in Section 2.5.2). We believe this is due to the unequal amplitude variances that result from first averaging across trials within person. Thus, the first recommendation we make is to use averages when analyzing such studies, and to be aware of the potential for false positives when reading studies with unbalanced designs that make use of the maximum.

Our second major finding is that the choice of window location is critically important when the conditions have unmodeled latency differences—that is, when the analyses are concerned only with peak amplitude, but there is a systematic difference in the timing of the component across groups or experimental conditions. Type I and Type II error rates increase as the window moves away from ideal placement, and the error rates increase more rapidly when there is smaller variability in latency across subjects. The maximum is more robust to window choice, and increasing the width of the window can also help capture the peak amplitude of both conditions in some cases. These findings lead us to suggest using the maximum when unmodeled latency differences may be present, and to caution researchers to check visualizations for this potential confounder.

A third major finding is that the simulations suggest the standard distributional assumptions are reasonable and for the most part hold in the simulated data. One

cause for concern is the IID assumption may be violated. One way to address this issue is to model the error structure direction using an approach discussed later in this section. This framework allows one to test whether an independent error structure holds in the data and compare fit of an independent error structure to more general error structures.

Overall, we find that each summary metric has strengths and weaknesses. Anticipated features in the data, the costliness of certain errors, or the desired interpretation may need to be stated to justify the choice of summary metric. This conclusion is frustratingly nonperscriptive, particularly for a problem where consistent methodology recommendations may be needed for replicating studies. But when one uses multiple metrics such as fit to distribution, Type I error rate, and Type II error rates, one procedure may not always dominate another procedure across all conditions for all metrics that were considered. That is indeed the situation here. Further, we note that the Type I error rates are at times quite discrepant from their nominal values. Usually, effective Type I error rates are slightly off from the nominal levels, and authors of simulations report problems when effective Type I error rates hit .10 or so. However, these simulations show extreme Type I error rates that at times reach .90 or greater.

The results of these simulations suggest that we need a substantial reconceptualization of the pipeline for ERP analysis. It may not be productive to limit our analysis to simple summary measures like the maximum or the average. Instead, the solution may be to develop a new analytic approach.

The current steps, outlined in Section 2.2, involve a carefully ordered process of filtering, averaging over trials, taking the summary metric, and then conducting statistical tests. As we have discussed, the order of steps 3 and 4 can impact the distribution of values when working with the maximum (but not the average). Indeed, many of the substeps involve nonlinear operations and are thus not exchangeable. It

is well-established [24] that the average waveform, or curve, may not be equal to the waveforms of the averages. A new analytic approach would be most useful if it avoided these issues of order of operations.

One option for leveraging the hierarchy of sources of variability in ERP studies and exploiting the well-studied shapes of components is to use nonlinear mixed-effects models. While Equation 2.5 is used as a data-generating function for the simulations in this paper, it could be used as an analytic model.

By using such a framework, we no longer need to be concerned with the order of operations during analysis. Instead, the entire process can be conducted simultaneously. It is not necessary to first average over trials—we can instead specify the hierarchy of the data (for example, trials are nested within conditions, which are nested within individuals; or trials are nested within individuals, which are nested within conditions) and the model-fitting procedure can accommodate an appropriate weighting and error structure. Thus, issues of unbalanced designs (whether from the experimental design or as a result of artifact rejection) are no longer a concern. Further, this framework provides a natural way to implement, estimate and test different error structures beyond IID.

We also circumvent the need to choose whether the maximum or the average is a more appropriate summary metric for a given study. Instead, we make use of the shape of a component from which we can derive different quantifications of the waveform while controlling for the many other sources of variability. For example, to compare latencies for a given component across groups or conditions, we can fit Equation 2.5 as a nonlinear mixed-effects model, constraining L to be near the anticipated time of the component (such as .300 for a P300). L is thus treated as a fixed-effects term, and the remaining parameters are controlled for as we test L directly.

However, such major changes to an accepted analytic pipeline require additional research that is beyond the scope of this paper. By pursuing this proposed new

approach, we can leverage recent advances in nonlinear mixed-effects modeling, along with advances in data science approaches, to provide a more unified and systematic foundation for the analysis of ERP data.

CHAPTER III

Basis Sets for Testing Meaningful Landmarks in Time Series Data

3.1 Introduction

The examination of trajectories, or curves in repeated measures data, presents several challenges for data analysis. Researchers frequently have specific hypotheses about properties of trajectories such as the time at which a peak amplitude occurs over the time series, whether the outcome variable returns to baseline after a stressor, or whether the asymptote of the outcome variable differs across two experimental groups or conditions. Such properties can be examined directly by positing specific functional forms whose parameters map onto these properties. There are several benefits of such an approach, as we will review in this paper, including (1) parameter estimation connects directly with the hypothesized properties of the trajectories, (2) heterogeneity of these properties can be addressed using random effect models and (3) the hypothesized properties can be tested across groups or conditions.

Traditional general linear model approaches to modeling trajectories, such as polynomial regression, may not always provide direct tests of such properties. A significant quadratic effect, for example, may not translate easily into a test of whether the asymptote of the trajectory differs across two groups or whether the timing for the

peak amplitude of the trajectory differs across two experimental conditions. Thus polynomial regression may be useful for general curve fitting but may not be useful for testing specific properties of those curves.

In this paper we discuss relatively simple methods by which hypotheses about properties of trajectories can be tested. The general idea is that one can use standard modeling approaches, such as general linear models or generalized linear mixed models, with predictors carefully selected so that the parameters of the model provide direct tests of the research hypotheses. In an analogous way to how sets of orthogonal contrasts can provide specific tests of research questions in an analysis of variance (e.g., do the means of these two groups differ from the means of those two groups?), the approach proposed in this paper uses a set of well-chosen predictor variables that parameterize a given property (e.g., do the peak amplitudes differ across these two groups?).

We use the concept of a basis set, which provides a way to fit data using simple linear models. There are several common basis sets in use, including orthogonal sets of contrasts in experimental design [30], sets of polynomial codes in regression [53], Fourier transforms [44], and particular types of spline models [43]. These basis sets possess well-studied properties that are suited to the mechanistic attributes of a particular application or field. Basis set parameterization can be used to estimate the functional relation between variables in a given application—the challenge lies in understanding the properties of the candidate functions and selecting the appropriate parameterization for the research question. Ideally, one can choose a basis set so that the parameters of the statistical model provide direct tests of the research hypotheses. For example, a set of predictors could be chosen so that each predictor corresponds to a property of the trajectory and then the coefficient associated with that predictor provides a direct test of that property. In this way we can select basis sets that are more directly related to the research question and provide justification for why other

basis sets may not be relevant.

Sometimes such a parameterization by well-chosen predictors is not possible. In those cases it may be possible to formulate the problem in terms of a nonlinear mixed model in which the parameters relate to the properties of the trajectories. For example, if one is interested in modeling a process that (1) increases from baseline levels at a particular time point, where that time point is estimated from the data; (2) increases at a rate that is estimated by the data; and (3) reaches an asymptote that is also estimated by the data, then a particular nonlinear parametric form with three parameters corresponding to those three properties can could be tested. Those parameters and their standard errors can be estimated using nonlinear regression (see, for example, Gonzalez & Wu, 1999) and can include random effect terms to model heterogeneity.

The methods outlined in this paper can be valuable to researchers working with physiological time series data such as EEG, fMRI, MEG, EKG, pupillometry, and other biological variables as well as data from wearables such as activity monitors and data from experienced sampling methods. The proposed methods are also useful for detecting outliers (not just points but also trajectories that are outliers relative to other trajectories), modeling variance and covariance structure over time and across variables, understanding statistical significance of trajectory parameters, having more efficient estimation procedures of trajectories and their properties, and inspiring further analyses and future studies.

The outline of the paper is as follows. First, we review the concept of basis sets and their use in linear models. Second, we focus on interpretability. This includes a discussion of reparameterization and an example to show how small changes in a basis set can lead to different conclusions even though the overall fits are essentially indistinguishable. Third, we provide an example of using the normal kernel as a basis set for ERP data. Fourth, we review regression spline mixed models and use that

model to extend the normal kernel model in the previous section. The paper ends with a summary and a set of recommendations for the applied researcher.

3.2 Basis Sets

In the model-building context, a basis set provides a mechanism to fit functional forms. Disciplines have conventional basis sets, e.g., polynomials are common in some areas of psychology and Fourier transforms are often used in engineering and physics. Spline functions offer additional possibilities for basis sets and each of these can be used to impose expected structure on a model, such as combinations of normal kernels or cubic functions. Some basis sets may seem contrived but can be useful in the appropriate setting, such as approximating the shape of an overall trajectory, and some basis sets can provide direct tests of hypotheses of interest. In order to reach a wide audience and increase the accessibility of the paper, we do not delve into the specific technical details of the properties of a basis set, extensions to linearly dependent vectors called frames, and other relevant details. Our approach, instead, is to provide basic intuition and examples of the usefulness of basis sets as a way to test hypotheses about trajectories.

3.2.1 Least Squares Review and Basis Sets

The least squares problem can be formulated as in Equation 3.1, where $x_i \in \mathbb{R}$ are fixed scalar values measured for each individual and the objective is to find the best coefficient vector β that minimizes the squared residual.

$$\arg\min_{\beta} \sum_{i=1}^{n} r_i^2 \tag{3.1}$$
$$\text{where } r_i = y_i - f(x_i, \beta).$$

Often, we have many measurements per person, so $x_i \in \mathbb{R}^p$ is instead a row vector

in a data matrix. In general, the function f can be characterized in matrix notation as $f(X, \beta) = X\beta$. We usually think of X as a matrix of predictors but the formulation is general and can include functional forms. For example, in the context of a time variable, t, we can place functions of t as columns in matrix X (the predictors), such as a linear function of t as one vector, a square of t to model the quadratic portion as a second, and so on, as in the case of the polynomials. We can also include the unit vector as a column in X to model the intercept. In symbols, the regression function can be modified while retaining the assumption that $f(\cdot)$ is linear in β:

$$f(t, \beta) = \sum_{j=1}^{q} \beta_j \phi_j(t), \tag{3.2}$$

where $\phi_j(t)$ is any function of t. If $q = 3$, $\phi_1(t)$ is a constant, $\phi_2(t)$ is the identity and $\phi_3(t)$ is the square, then this becomes the usual quadratic regression formulation.

However, the polynomial basis is not the only basis one can use. It is possible to reformulate the design matrix X to use any functions ϕ_j. For example, using sines and cosines of t still yields Equation 3.2 as a linear combination with coefficients β_j interpreted as the weights associated with each ϕ_j, or basis vector. This gives rise to a Fourier model.

Another representation included in this framework is the Taylor series expansion, which is a linear combination of increasing orders of derivatives. If one takes, for example, a linear combination of first and second derivatives at a particular point along the trajectory, then the associated parameters provide tests of change and acceleration, two concepts that may be relevant in some applications such as developmental psychology.

Yet one more example of a different basis set involves linear transformations of an existing basis set, such as a rotation, or a change of coordinates. Some psychological phenomena may be expressed more naturally in either a rotated space (e.g., the psychological experience of ambivalence) or in a different coordinate space such as

polar coordinates.

These examples show that functional relationships between variables can be approximated by linear combinations of component functions. If the component functions are well-chosen to represent specific hypothesized properties of the trajectories, then the resulting parameters estimate those properties and their standard errors allow computation of confidence intervals and statistical tests. However, not all properties of a trajectory can be modeled simply as a linear combination of well-chosen basis sets as in Equation 3.2. In those case, a nonlinear model may be appropriate. One could create a linear approximation (e.g., via a Taylor series expansion), or one could turn to nonlinear regression. Our recommendation is that the analyst choose the representation that leads to interpretable parameters that are connected to the original hypotheses.

3.3 Interpretability of Coefficients

A major problem for the representation in Equation 3.2 is that for some basis sets the coefficients β_j may not have direct interpretation to psychological phenomena under study. As discussed in the next section, this problem was addressed via reparameterization by Cudeck and du Toit [13] in the case of the quadratic polynomial. While the unit vector, linear t and quadratic t^2 are commonly used in time series analysis as the polynomial basis set mentioned earlier, the coefficients β_j corresponding to the polynomial basis set may not have direct interpretation. For example, it may be difficult to attach psychological meaning directly to the β_j for the quadratic term other than the usual regression interpretation of the contribution of the associated vector controlling for the linear combination of all other vectors in the model. What meaning would the quadratic β have in adjudicating between predictions of psychological models, especially when the psychological models are likely not expressed in terms of the β_j coefficient of t^2 having adjusted for all other terms?

The psychological theory may be expressed in a vague manner such as a statement about whether or not the trajectory deviates from linearity (e.g., is concave or convex). In such a case the hypothesis refers to the overall curvature of the trajectory and for these kinds of hypotheses the polynomial basis set may be sufficiently flexible to provide adequate fits to the observed trajectories. However, alternative basis sets may provide similar tests of curvature while also exploring potential updates to the existing theory.

There are also cases when psychologists are interested in testing specific properties of trajectories. For example, does the time at which the maximum occurs vary across two groups? does the rate of increase differ across two experimental conditions? does one group rebound more quickly than another group? The βs from the commonly used polynomial form of Equation 3.2 do not directly provide answers to such questions, without possibly reparameterization. In other words, the standard polynomial approach of modeling nonlinear trajectories may do well at capturing the overall shape of a curve, but may not be helpful at testing specific properties about that shape. When researchers have hypotheses about such specific properties, then the usefulness of the polynomial basis set as typically estimated becomes questionable.

3.3.1 Reparameterizing the Polynomial

Cudeck and du Toit [13] provide a way to salvage the polynomial approach in the case of testing specific hypotheses about trajectories. They show that one can transform the standard three βs in Equation 3.2 to yield three new coefficients (a reparameterization of the original coefficients) that have more direct interpretations to properties of the trajectories. They specifically transformed the three βs into three "landmarks" or properties of curves: at what Y value does the curve start (intercept)? how high does the curve go before reaching the peak (maximum, a value on the scale Y)? and at what time does the curve reach its maximum (maximizer, a value on

the scale X)? That is, the values of the three coefficients are directly interpreted as "start", "maximum," and "maximizer." The coefficients of the model estimate those quantities. Thus, any test of significance of those rescaled parameters directly tests, for example, group differences between the parameter values. Both the original model formulation and the new model have identical number of parameters (three) and identical fits as they are the same model. One formulation provides a better interpretation of the polynomial basis set in terms of landmarks that can be tied directly into what the researcher is interested in testing.

Cudeck and du Toit [13] showed that the quadratic polynomial in Equation 3.3 can be reparameterized in the following way:

Begin with the quadratic polynomial:

$$g(\beta, x) = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{3.3}$$

We call this polynomial a basis set because it is a linear combination of the vectors $[1, x, x^2]$. When x equals zero, the intercept is

$$\alpha_0 = \beta_0.$$

Taking the derivative of $g(\cdot)$, setting it equal to zero, and solving yields the maximizer:

$$\alpha_x = \frac{-\beta_1}{2\beta_2}.$$

Plugging α_x in to $g(\cdot)$ yields the maximum:

$$\alpha_y = \beta_0 - \frac{\beta_1^2}{4\beta_2}.$$

Algebra yields $h(\cdot)$, which is equivalent to $g(\cdot)$ when using the αs as described

above.

$$h(\alpha, x) = \alpha_y - (\alpha_y - \alpha_0)(\frac{x}{\alpha_x} - 1)^2. \tag{3.4}$$

This new formulation $h(\cdot)$ is now directly interpretable in terms of the coefficients—the estimated intercept, maximum, and maximizer. These estimates are based on the quadratic form that we began with.

One could use the delta rule to approximate the standard errors of these new parameters but Cudeck and du Toit [13] chose to reformulate the problem into a nonlinear regression problem so that the three parameters (start, maximum, and maximizer) and their standard errors are directly estimated from a nonlinear regression model. Preacher, Hancock, Harring, and Hancock [50] extended this approach to other functional forms and other hypothesized properties about the trajectories by linearizing the functional form and then using the derivatives as constraints in a structural equation modeling approach.

While the nonlinear regression framework generally allows for added flexibility in the model parameterization, the key difference in testing this model using the original polynomial basis set as opposed to the nonlinear regression framework proposed by Cudeck and du Toit [13] is the way the error (i.e., the "plus ϵ" part of the model) is characterized. In the original polynomial basis set version the error term is additive in the polynomial, but in Cudeck and du Toit [13] the error term is additive in the nonlinear regression version.

3.3.2 Using Different Basis Sets

The polynomial basis set implemented in Equation 3.2 is only one of many possible basis sets one can use. It may appear that we could select any basis set to approximate the shape of the curve, then follow the above steps to estimate the coefficients of the linear model. However, it turns out that different choices for the basis set can have

substantial effects on the conclusions about the hypothesized "landmarks." This is because the choice of basis set function(s) ultimately equates to a set of shape assumptions. For example, the quadratic polynomial is symmetric and allows for both increases and declines over values of x.

A different basis set may make different shape assumptions. To illustrate, we have selected another basis set that yields a similar shape, but without the symmetry constraint.

$$k(\beta, x) = \beta_0 + \beta_1 \sin(x) + \beta_2 e^x \tag{3.5}$$

This basis set, which is a linear combination of the vectors $[1, \sin(x), \exp(x)]$, combines the sinusoidal curve with the increasing slope of an exponential function. In practice, this basis set can model arch-like curves in the data that appear to be pulled to the left or right instead of being symmetric. Since we anticipate only one peak in the data, and not periodicity, we rescale x to fall in the range of 0 to π.

This example uses data from Smith and Cook [58] for daily levels of serum-creatinine for a renal transplant patient (used by Cudeck and du Toit [13]). Increasing levels of this substance indicate that the kidney is functioning normally, while decreasing levels suggest a rejection of the new kidney. Thus the changepoint, or maximum, and the day when it occurs are of primary interest to researchers in this area.

We fit both the quadratic polynomial from Equation 3.3 and the sine-exponential from Equation 3.5 to this dataset.

The standard quadratic polynomial in black ($R^2 = .92$) and our alternative sine-exponential form in red ($R^2 = .93$) both fit well (and if anything the alternative is preferable due to its higher R^2), but yield practically different maxima and maximizers. The quadratic finds a changepoint at 5 days, whereas the sine-exponential finds it at 6 days. This difference could make a dramatic difference for patients. Of course, one could put confidence intervals around these estimates and test for differences;

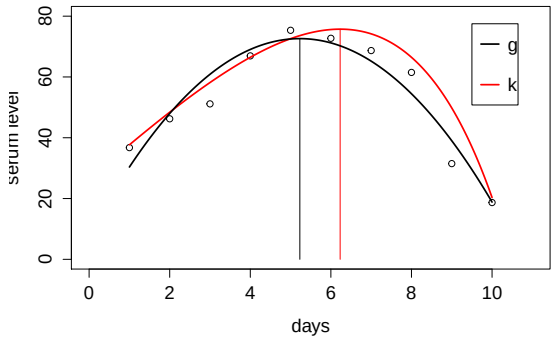


Figure 3.1: Fitting two different data transformations

given sufficient power these estimates may statistically differ.

Our point in this subsection is not to advocate one basis set over another but rather to highlight that one's choice of a basis set can lead to dramatic differences in the interpretation and conclusions. As we have shown, this can happen even in settings where overall curve fit is relatively high and similar across competing basis sets. Thus, we argue that researchers should consider both their research question and known properties of their data when selecting a model.

The basis set approach can be used to test specific properties, for example, the location of the maximizer on the scale X. It can also be extended to test specific consequences of basis sets. Examining such implications of basis set choices on predictions of key curve properties and landmarks can provide directions for new study designs, suggest new data to collect, and provide insight into new tests focusing on specific constraints.

3.4 An Event-Related Potential Example

We have stated that it is common for psychological researchers to test hypotheses about properties of curves, such as the difference in the maximum across two groups or two experimental conditions. An example occurs in the domain of event-related potentials (ERPs), which is a widely-used and non-invasive way to quantify the brain's

response to a variety of stimuli. A typical ERP study seeks to identify voltage differences in a particular portion of the ERP waveform across groups or experimental conditions. After computing an average waveform for each subject within conditions [42], an ANOVA is used to investigate the average or maximum voltage near a prespecified time across different groups or conditions. In this way a property of a curve (the maximum) is tested across two groups or conditions. However, this approach simplifies data in ways that are no longer necessary and it turns out makes assumptions about the underlying functional form that may affect the conclusions in major ways. We offer a general analytic framework that provides alternative approaches and contains the existing approach as a special case.

In ERP research, there are a few standard questions. They include: "Is there an amplitude difference in the peaks of two groups near a certain time?" and "Is there a latency difference in the peaks of two groups for a particular component?" However, our framework allows for more complex questions about patterns in the ERP, such as: "Is there a difference in the relationship between amplitudes of pairs of adjacent components across groups?" and "Do trajectory patterns differ across different channels?"

Basis sets provide a means for fitting a single waveform to the ERP data. Demographic information, experimental conditions, and the natural hierarchy of the experimental design (including both repeated trials and multiple channels, or recording sites on the scalp) can be included in the context of mixed-effects modeling that allows for individual differences in the parameter values (e.g., the value of the maximum). Embedding the basis set approach in a mixed-effect model has other advantages such as being able to model more general error structures over time (e.g., Preacher & Hancock, 2012).

Mixed-effects models (also known as multilevel models; varying-intercept, varying-slope models; random effect; or hierarchical models) take into account the variation

between groups. These models include fixed effects that do not change across groups and random effects that do. These models can include additional variables, such as subject-level age, gender, or disease status.

More formally, we can write:

$$y_i = X_i\theta + Z_i\gamma_i + \epsilon_i \tag{3.6}$$

for the outcome y_i of groups i where X_i and Z_i are a design matrices for fixed and random effects, respectively [28]. For example, to model how voltage on a single channel differs in two conditions, we might model the average voltage for condition i as

$$V_i = \theta + \gamma_i, \tag{3.7}$$

where θ is the voltage contribution from the overall study and γ_i is the remaining portion of the voltage for the specific condition.

In this example data, a subset from Begleiter [4], we make use of 1 second of data for 40 subjects' P7 waveforms over approximately 50 trials for each of 2 conditions. This dataset is on a much smaller scale than ERP researchers typically examine, but it demonstrates how complicated data from a typical ERP study can be. The data are plotted in Figure 3.2, where it is difficult to see any clear patterns due to the density of the data. Even in spots where specific patterns are identifiable, there are differences in amplitudes and latencies across channels.

Our basis function is the kernel of a normal function:

$$s(t, h, m, v) = h\mathrm{e}^{-\frac{(t-m)^2}{2v^2}}. \tag{3.8}$$

By using the functional form of a normal distribution kernel as a basis set, the coefficients for fixed and random effects become interpretable as the location of the

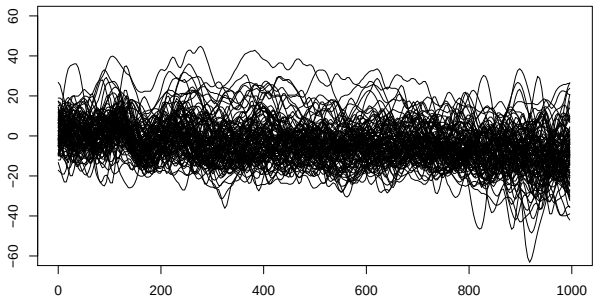


Figure 3.2: Plot of all trial-level waveforms

component, its width, and its height. This is particularly useful when both the time latency (such as 100 milliseconds for a P100) and amplitude (or voltage) are of interest. It is standard to estimate the coefficients with standard errors, meaning that confidence bands can be constructed and hypothesis tests performed for these values. Many hypothesis tests currently performed in ERP literature involve voltage differences between groups, for an event that happens at an approximate time, so using normal kernels seems immediately useful.

Since the selected functional form is not linear in the parameters, we fit this normal kernel basis set using a nonlinear mixed model:

$$y_i = s(\phi_i, t_i) + \epsilon_i, \tag{3.9}$$

$$\text{where } \phi_i = A\beta + B_i b_i. \tag{3.10}$$

The parameter vector ϕ organizes the coefficients (which are contained in β and b_i) to reflect the specified hierarchy in the model. Here, the vector of voltages for condition i over all times t is estimated as follows:

$$\hat{y}_i = (h + h_i)\mathrm{e}^{-\frac{(t-(m+m_i))^2}{2(v+v_i)^2}}. \tag{3.11}$$

The "mean" parameter, m_i, controls the center of the component on the time axis for each group. The "variance" parameter v_i, which controls the width of each com-

ponent wave, is also estimated per group. The coefficients h and h_i are estimated as fixed and random effects to capture the variability in amplitude in groups, relative to an average-fit waveform. Thus, in the mixed-effects model, the fixed effects determine average waveform while the random effects determine amplitudes and latencies per group. For this example, we are interested in comparing the amplitude of a component near 250 milliseconds post-stimulus across the two conditions, so we constrain m_1 and m_2 to be near 250 milliseconds. The resulting fit is easily visualized with ± 2 standard error bars at the group level in Figure 3.3.

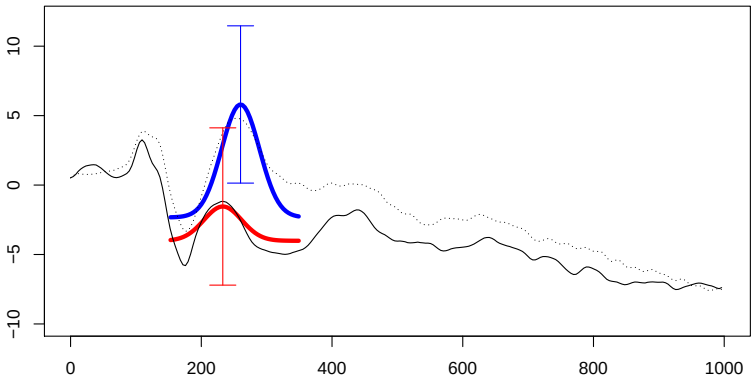


Figure 3.3: Comparison of two conditions using normal kernels with error bars

The original question asked by the authors was: "Is there a voltage difference across conditions at 250 milliseconds?" Here we can see that there appears to be a difference, but that it turns out to not be statistically significant. The location m_i can be considered a nuisance parameter in this case, but it could be of primary research interest to a researcher focused on latency differences. One could even formulate hypotheses about the joint distribution of the amplitude parameter and the latency parameter, given that there may be interdependent group differences on both of these parameters. Existing methods do not easily provide examination of hypotheses on both of these parameters. This example illustrates how we can make use of a new basis set to capture different properties of these waveforms.

3.5 Regression Spline Mixed Models

One possible extension of the normal kernel model we just introduced involves regression spline mixed models. Regression splines are a nonparametric option for fitting waveform data. Often, the term spline is used to imply polynomial splines. For these, landmark points, such as local maxima and minima, make natural choices for knots. Smoothing splines, also based on polynomial functions, are occasionally used for EEG data [32], but resulting waveform estimates vary substantially based on choice of tuning parameter and can be difficult to interpret or test formally. This paper instead suggests use of generalized splines, which consist of piecewise functions of any form. Whereas the previous example used only one basis set function, this framework allows researchers to specify as many or few basis functions as are needed for their hypotheses. The regression spline model is expressed as:

$$Q = \sum_{k=1}^{r} c_k s_k(t) \tag{3.12}$$

where $s_k(\cdot)$ are basis functions of any type and $c = (c_1, ..., c_r)$ are unknown coefficients to be estimated [47]. The functional forms of $s_k(\cdot)$ can again be defined so that the fitted coefficients c_k and parameters ϕ have meaningful interpretations to researchers, much like in Equation 3.2 and the previous section. This is a key idea from functional data analysis [52], but the gap currently lies in determining the functional form. Our basis sets bridge this gap and provide the means to test a wide variety of scientific hypotheses. While there is a connection with the basis set approach we take and the usual one found in functional data analysis, the former relies on differential equations and supporting material such as phase plots. This makes our approach perhaps more accessible to behavioral and social science researchers.

We also bring the hierarchical structure of the data into the model via mixed effects modeling, as done in the previous section. Depending on modeling goals,

we can allow regression splines to have either fixed or random contributions in the model, or both. To combine mixed-effects models with regression splines, the time series measurement of an individual subject can be written as the sum of a (fixed) population mean function and a random function (both estimated nonparametrically) along with white noise. Thus, we get a model for the overall fit, but also gain insights about more granular levels in the data. Being able to see how each channel compares to the average fit gives us a good impression of the variability in the model, and which channels are more or less similar. This also helps us to deal with unbalanced designs. It is important to visually inspect the model fit at each level, since it is possible that there are anomalies in the data that are not apparent when plotting only the fitted curve.

Mixed-effects procedures such as this one are implemented in any standard statistical software, such as the `nlme` function in `R` by Pinheiro, Bates, DebRoy, Sarkar, and R Core Team [49]. This function is based on work by Lindstrom and Bates [39], with updates to allow for nested random effects and more complicated error structures. Regression spline mixed models have previously been studied by, for example, Mackenzie et al. [43], to explore nonlinear longitudinal data.

3.5.1 ERP Combination Example

In practice, EEG datasets can be rather large. They may span long periods of time. Even short ERP datasets still involve 20 or more people, for whom up to 345 channels can be collected during each of many trials of several experimental conditions. Often, researchers seek a smoothed representation of the EEG dataset. In this section, we will show how regression spline mixed models (RSMM) can combine the features of splines with a hierarchical random effects framework that goes beyond simple smoothing to explore EEG data at any of the many levels that are collected and of interest to researchers.

One aspect that an overall spline fit does not account for is that the individual channels may be of interest in some cases. For example, we know that brain functioning is local, not global—so we might want to examine separate regions or channels and compare them. We can also check for malfunctioning sensors by visually inspecting or formally testing for abnormal readings in single channels. If we do remove a malfunctioning sensor through this or other methods, it is unclear if using an averaged fit via existing methods is still accurate, or if the remaining sensors should be reweighted to reflect a balanced contribution of all brain regions. A mixed-effects modeling approach provides one way to address such unbalanced designs.

We utilized the `nlme` package in `R` [49], which will fit coefficients for any functional form as both fixed and random effects. This approach currently requires the user to prespecify the number of peaks to estimate. This is where the ability to translate a verbalized hypothesis into a basis set becomes important. A scientific rationale should exist behind the design of the basis set, just as one should select a hypothesis and decide what results will be considered significant before performing any statistical test. This approach can be used for both data exploration (as shown in Figure 3.5) and testing.

To illustrate the utility of implementing this situationally-appropriate basis set in RSMM, we use another subset of the EEG dataset from Henri Begleiter at the Neurodynamics Laboratory at the State University of New York Health Center at Brooklyn, made available through the UCI repository [4]. Shown in Figure 3.4 are 64 channels of EEG data for a single ERP trial of a single subject.

To display the data, neuroscientists typically want a plot that is smooth and shows only a couple noticeable oscillations in the data over the entire time period. This averaging takes place over several trials and several people, without accounting for these differences in amplitudes (beyond baselining) and latencies, and the variability in both increases drastically at this scale (glossing over various preprocessing steps

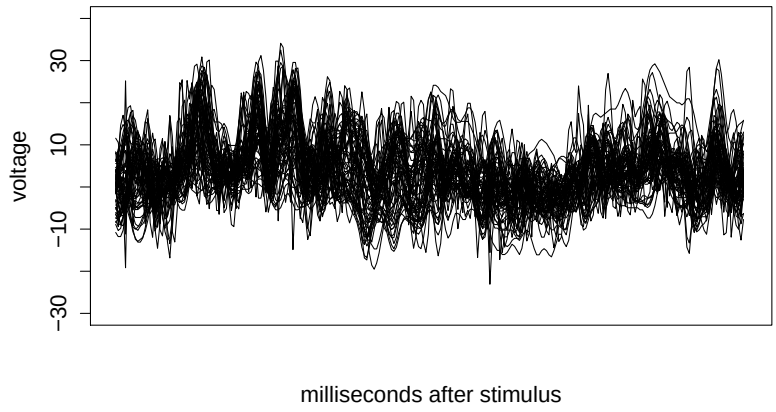


Figure 3.4: ERP data for 64 channels of 1 trial

that are typically done). To mimic this goal of exploring the ERP and providing a general model, we have fit a mixed-effects model, this time using several normal kernels over the longer window of interest, where fixed effects determine latencies and an average waveform and random effects for each channel determine separate amplitudes, shown in Figure 3.5.

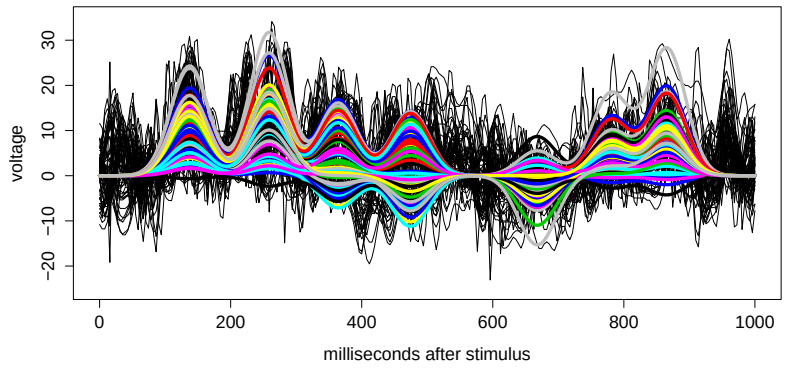


Figure 3.5: Mixed-Effects model using normal kernels on ERP data

This model is still an over-simplification of the data. While it does account for, and even showcase, the variability in amplitude across channels, the variability in latencies has not been included in the simple toy model. The model in Figure 3.5 was fit using a basis set consisting of 7 normal distribution kernels, where the "mean"

parameter (which controls the center of the component on the time axis) is selected as a fixed effect in the model-fitting process using REML, the "variance" parameter (which would control the width of each component wave) is fixed, and coefficients are estimated as random effects to show the variability in amplitude among channels. The plot shows that there is high variability in amplitude across these 64 channels at these 7 temporal locations.

The voltage v_p at each timepoint within a channel is estimated with normal kernels in the RSMM framework as follows:

$$\begin{aligned} v_p = \beta_0 &+ \beta_1 \mathrm{e}^{-600(t_p - m_1)^2} \\ &+ \beta_2 \mathrm{e}^{-600(t_p - m_2)^2} \\ &\vdots \\ &+ \beta_k \mathrm{e}^{-600(t_p - m_k)^2}. \end{aligned} \tag{3.13}$$

In this example the parameter k was set to 7, chosen by visual inspection and model complexity concerns. However, the number of normal kernels could be set by the design, such as a kernel for each time a stimulus is presented, or could be driven by the subject, such as normal kernel every time a response is initiated. Kernels could also be constrained to the number of anticipated components, such as 2 for the N200 and P300. Computation time increases approximately exponentially with k, which might be an important practical concern for certain applications. In the future, k could be a free parameter selected during the model-fitting process. This would be useful in model-building, rather than hypothesis-testing, settings. It could be useful to see where components are identified by the algorithm to minimize error, and how this compares to visual inspection of the data. This could inspire further investigation and formalized hypotheses that might not otherwise be explored. In these cases, k should still be constrained in some reasonable way to avoid overfitting (as $k = i$ would lead to a perfect fit).

The variance parameter of 1/1200 was chosen to match an existing `R` package (`eegkit` by Helwig [32]), which includes a function for simulating EEG data using this value based on cross-validation over many datasets. This value was also selected by EEG researchers in informal discussions as visually appearing the most like an ERP component among several options. In the future, allowing this value to vary may be informative, but will drastically increase computation time.

Depending on the problem, it may also make sense to allow the mean parameter(s) to vary. There is variability in the latency, or response time, across both channels and people. For example, individuals with depression have much slower response times, and so they are sometimes removed from analysis. By allowing the random effect for groups of people to vary, we can include these subjects in the analysis. This would lend more power to statistical tests and could help neuroscientists explore and better understand neurological implications.

The present numerical approach is sensitive to the choice of start points. Fortunately, visual inspection usually gives a good impression of whether the model has been misspecified, and how to update these values before re-running the algorithm. This does require some finesse, but becomes straightforward with experience. Future work may involve overcoming this sensitivity by repeated fitting, establishing best practices for selecting initialization values or using different approaches such expectation maximization or Bayesian models.

3.6 Conclusions

Future work involves an in-depth comparison of existing and proposed methods, particularly with respect to standard error calculation and testing. While there is an argument that interpretability can be improved by selecting an appropriate basis set, a statistical justification would be valuable. By performing analyses inside a mixed-effects framework, we believe that individuals, groups, and trials that would otherwise

be excluded can be modeled through additional random effects terms. This should increase our effective sample size, and thus decrease standard errors and increase power. However, this also decreases the degrees of freedom for analysis. These degrees of freedom may be compensated by a well-chosen basis set that could require fewer terms to achieve the same prediction error rates.

Creative use of basis sets may inspire new approaches to other statistical issues, for example, using a polynomial basis set to form a general framework for higher-order interaction models. Traditional approaches to testing continuous-by-continuous interaction models involve artificially grouping one variable and exploring how groups moderate the relationship between the continuous predictor and response. Further, interactions are usually limited to "linear by linear" interaction terms, and rarely consider interaction of higher order terms such as "linear by quadratic" or "quadratic by quadratic." The polynomial basis set provides one way to explore higher-order interactions but exploration of alternative basis sets may give rise to new approaches to model interactions.

Currently, covariance structures are assumed to be relatively simple. However, autocorrelation among observations is expected in time-series data. Exploration of how best to include additional assumptions on the errors at the many levels of the hierarchy will be valuable and is suggested as future work in many RSMM papers (such as Lindstrom and Bates [39], Rice and Wu [53]). In the ERP context, the covariance structure is key to building a more complete picture of brain activity. We can use it to explore the spatial relationships across scalp electrodes and how signals might be traveling and dissipating across the scalp.

Simulations, theoretical derivations, and applied analysis will all be important steps in validating this method. A comparison of methods, addressing model mis-specification, prediction error, and computation time, should be explored in future work. Since the interpretability of parameters is such an important aspect of this

method, it will also be important to qualitatively compare the performance of this method to existing approaches in context. Ease of use, ease to communicate the method, and consistency of scientific conclusions should also be considered.

This methodology draws inspirations from many fields. We have mentioned direct ties to nonlinear regression, functional data analysis, analysis of variance contrasts, and linear regression. However, related ideas also exist in kernel density estimation, nonparametric statistics, link functions, and other areas. As we bring these methods together, many theoretical properties should naturally follow and facilitate comparison to existing approaches.

CHAPTER IV

Applet-Based Training for Identifying Appropriate Statistical Methods

4.1 Introduction

Technology in the classroom has long been a subject of study [54, 63]. Research generally shows that technology is positively linked to student performance, particularly when the technology supplements classroom instruction. In statistics, this is often implemented as simulations and demonstration applets [8], such as those which demonstrate Simpson's paradox [56], probability distributions [36], power [1], the central limit theorem [18] and others [62]. These tools help students to get a better feel for quantitative properties of statistical concepts. Repeated meta-analyses have shown that technology use has a meaningful impact on statistical learning [33, 55, 59]. We have implemented an applet which allows students to practice statistical problem-solving in real-world contexts, rather than exploring quantitative properties.

Students in an introductory statistics class generally learn the details of a variety of statistical methods. They learn how to construct confidence intervals and conduct hypothesis tests to make inferences for many different population parameters. Often,

the focus is on how to carry out and interpret the various calculations after the parameter of interest has been identified—for example, homework questions may be given with the heading of population proportions, signaling a topic area to students. We noticed that a common difficulty for many students is deciding which set of methods is most appropriate to use in a given scenario. Lovett and Greenhouse [40] point out that learned knowledge tends to be context-specific and students can experience failure to transfer once they are not aware of the problems context. In introductory statistics, it is often the case that students learn in a context where they are told which test to perform, but are then assessed without that scaffolding. An applet called Name That Scenario (sometimes abbreviated NTS) was created to provide students that needed guidance and practice on how to identify the appropriate statistical methods to address a given research question.

This applet aligns well with a number of the new core GAISE recommendations, namely: to teach statistical thinking as an investigative process of problem-solving and decision-making, to focus on conceptual understanding, to use technology to explore concepts and analyze data, and to use assessments to improve and evaluate student learning [27]. NTS was designed with ease of use and student-centered learning in mind, in keeping with the overview of technologys use in statistical education by Chance, Ben-Zvi, Garfield, and Medina [7]. We believe that following these principles allowed us to create a learning tool which promoted motivation to persist in practicing and mastering this challenging skill. By showing students that they are making progress towards mastery, we encouraged a shift from a fixed mindset (wherein students gave up upon deciding that they were not good at this skill and would never be able to do this) to a growth mindset (wherein students believed that they could improve their performance through practice) as proposed by Dweck [21].

Our primary hypothesis was that use of NTS would improve student learning of this important skill, as measured by assessments throughout the term. A study was

conducted to test this hypothesis.

4.2 Applet Development

In an introductory statistics course, students are generally exposed to a variety of statistical procedures. One important skill is to determine which statistical methods or procedures are appropriate to address the research problem or question of interest. In the past, our students practiced this skill and were assessed through matching exercises given on paper.

As part of a new initiatives grant, graduate students were trained in finding, evaluating, and designing online learning objects. A statistics graduate student proposed the development of an online learning object to provide practice of the skill and it became Name That Scenario. As a research group, we saw value in providing this practice in an online format and created the first online version. The tool has since moved to a more stable platform that allows for data collection and analysis.

Name That Scenario is now housed online, in a course-specific learning platform. Students log in, enabling tracking of their progress in many aspects of the class including NTS. The design of the applet is quite simple. When students start the NTS applet, they first select at least 2 frameworks from which to receive practice. This enables the applet to be useful throughout the course, rather than only after all frameworks have been covered. They are then given a series of 10 scenarios from the selected frameworks and are asked to select the appropriate test to assess the given research question. Figure 4.1 shows this initial dashboard screen for selecting the desired frameworks. After each question, the tool provides problem-specific feedback, based on either a correct or incorrect response. Students receive a score out of 10 at the end of the 10 questions.

The NTS tool is readily available for students to use at any time in the semester. It is introduced after the first five statistical methods (one proportion, two proportions,

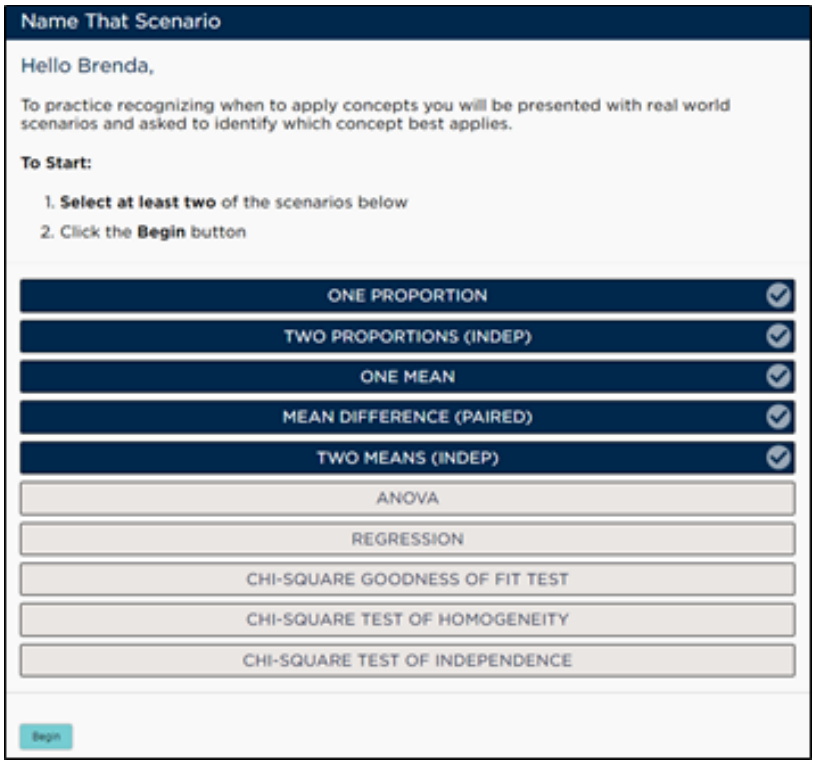


Figure 4.1: The landing page of the applet

one mean, paired means, and difference of two independent means) have been introduced in lecture. There are currently 250 total questions available across 10 types of statistical frameworks (those listed above, and ANOVA, regression, chi-square goodness of fit test, chi-square test homogeneity, and chi-square test of independence). This bank of questions continues to grow with regular updates.

This applet is a favorite of both students and instructors. Students generally highlight the approachable, convenient practice that NTS provides for a challenging aspect of the introductory statistics course. Instructors appreciate having a simple, calculation-free resource to recommend to students. The following key features of NTS are ones we consider universal to the success of any tool that gives students productive practice.

4.2.1 Providing Feedback for All Responses

In this way, NTS reinforces the thinking and learning process by providing a good model for students. Svinicki [61] discusses the value of such immediate feedback on student learning. The feedback was not difficult to program into our applet. We only require one feedback statement for correct responses which reiterates the important aspects, and a second for incorrect responses which largely emphasizes the same points. Figure 4.2 shows the interface of the applet for a correct response and an incorrect response. This feature could be expanded to include a different feedback statement for each student response option if there are important pairwise comparisons to make between particular detractor responses and the correct answer.

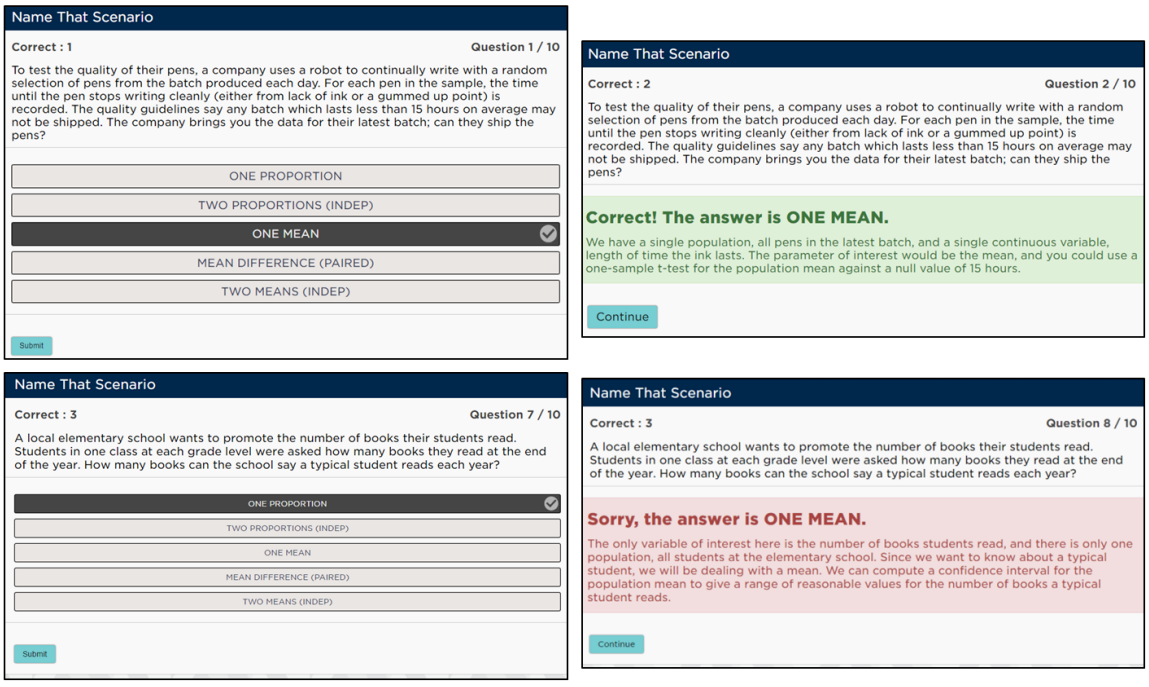


Figure 4.2: Feedback is given for both correct and incorrect responses

4.2.2 Summary of Performance

Upon completion of a set of 10 questions, students receive a summary of how many questions from each testing framework they were given and how many they got

correct within each. For each framework containing an incorrect response, the results page lists which test(s) they had incorrectly chosen. Thus, if a student consistently picks one proportion instead of two proportions, this common error would show up here. Figure 4.3 shows an example results page. This student correctly classified 3 out of 10 scenarios and made repeated errors by picking two independent means when the design was paired, and selecting two independent means when the variable of interest was categorical and thus the test should have been for two proportions.

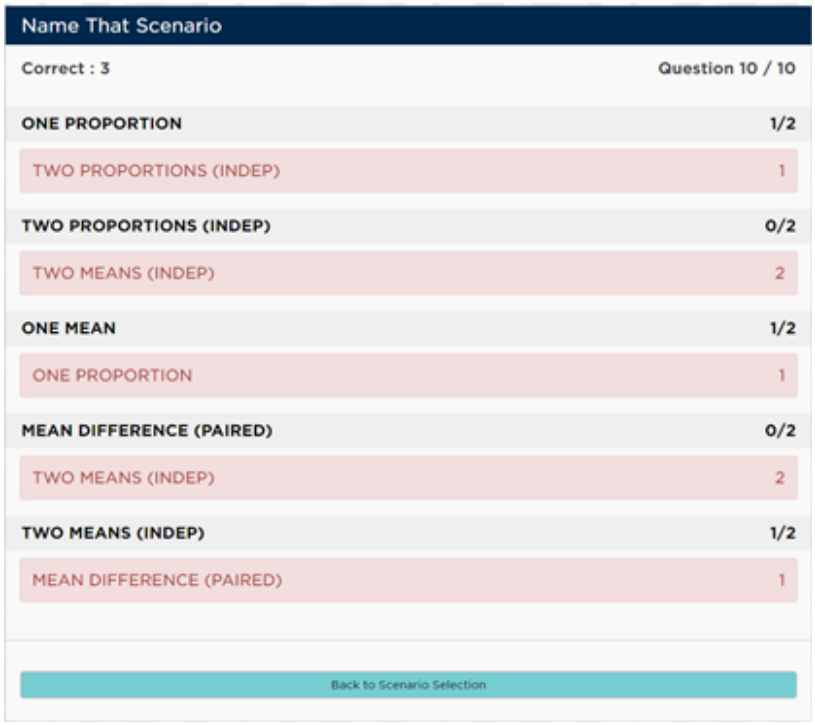


Figure 4.3: An example results page in NTS

In the future, we would like to use existing functionality in the learning platform to keep a running history over repeat visits to the applet. This could not only guide students to areas for improvement, but would also emphasize their progress. Eventually, this could be used to provide tailored clarifications and suggestions for further practice. We believe that providing students with this type of analytic information about their learning can inspire motivation to keep them working on challenging concepts.

4.2.3 Repeated Practice

Having a large bank of scenarios is one of the features that keeps students coming back to NTS. They generally see new questions upon repeat visits. Our ever-growing bank of scenarios is currently at 250 entries, with the earlier course topics oversampled. Hsu [33] and Schenker [55] both found that drill and practice style applets, such as ours, were some of the most effective computer-assisted instruction formats in statistics. One way in which we gather additional content is to have students create the scenarios themselves. Creation of new material can help students develop higher-order thinking about the framework selection process [5]. For example, Bates, Galloway, and McBride [3] showed that physics students who generated their own questions performed better on their final course examination. In the future, we may implement some form of gamification or badges to further encourage students to participate in this productive practice.

4.2.4 Portable, On-the-Go Practice

The GAISE report [27] states that "It is important to pick technology that does not become an additional burden for students or that hinders them further from meeting goals or objectives'' when using technology to meet GAISE standards. It also suggests that Interactive applets can be used to emphasize important statistical concepts without being encumbered by lots of calculations. By housing the applet online, students can use it anywhere they have internet access. The concepts being tested involve no calculations, so students can easily run through a set of 10 scenarios in 5-10 minutes without pen and paper. These features are some of the most highlighted by students and instructors alike.

4.3 Assessing the Efficacy of the Learning Tool

In the Winter term of 2016, a study was conducted to assess the effectiveness of NTS. Students enrolled in introductory statistics at a large, public research university in the Midwest followed standard curriculum. Once students had been introduced to five testing frameworks (one proportion, two proportions, one mean, difference of two independent means, and paired means), they were given the opportunity in class to take a pre-test administered through their course management system with the same structure as the NTS applet. The quiz contained the eight scenarios marked as pre-test items in Appendix A. These scenarios were selected and temporarily removed from the existing pool in NTS to be a representative sample of concepts across a range of difficulties. Appendix A also shows the multi-term cumulative percent of correct responses for each item to reflect item difficulty. When students completed the pre-test, they were allowed to see their responses and the correct answers. Immediately after the pre-test was completed, the NTS applet was made available to all students, regardless of whether they completed the pre-test. Students had unlimited access to NTS for the following week as class continued normally.

After one week with access to NTS, students were encouraged to take an in-class post-test. The scenarios, also shown in the Appendix, were selected to match the frameworks and difficulties of the questions given in the pre-test, and had been removed from the bank of questions at the start of the term. The concepts were given in the same order in each assessment, to avoid order effects. Again, students were shown the correct answers immediately after completing the post-test. A comparison of performance on the pre-test and post-test is the focus of our analyses.

Students were encouraged to use the applet as they prepared for their next exam, which took place approximately one week after the post-test was offered. The exam contained a Name That Scenario section, which tested students abilities to identify the correct statistical method. In our analyses, we present the average trajectories for

students who used the applet at various time points between the pre-test, post-test, and exam.

Of 1799 students enrolled in the course, 1025 completed both the pre-test and post-test. There was more material to be covered in the class held during the week of the post-test, including some exam preparation, which we believe led to the majority of missing post-test scores. However, we are confident that the remaining students who did participate represent the overall student population with respect to both gender (Table 4.1) and class rank (Table 4.2). Note that some students were in their second semester of their first year of college (freshmen), but had a higher class rank (sophomore) due to incoming credit hours. Thus, it is also possible that juniors by credit hours may actually be second-year college students, and so on.

	Female	Male
Overall Enrollment	0.5258	0.4742
Complete Data	0.5346	0.4654

Table 4.1: Self-reported gender of students

	Freshman	Sophomore	Junior	Senior
Overall Enrollment	0.1136	0.4744	0.3001	0.1119
Complete Data	0.1260	0.4746	0.2998	0.0996

Table 4.2: Class rank of students

We first compare pre-test and post-test scores for all students, regardless of NTS use. A simple paired t-test showed significant improvement in scores between the pre-test and post-test, $t(1024) = 9.14, p < .001$. During this time, scores for everyone improved by an average of approximately half a point (out of 8 points possible). Results support a statistically significant improvement in score, on average. The overall population mean improvement is estimated to be between .435 and .673 points with 95% confidence. This suggests that some improvement in score is due to in-class exposure to content or homework.

Figure 4.4 shows that the average accuracy on almost all framework types improved. Topics are ordered in terms of novelty in the legend and color scheme (newest last). The horizontal axis shows long-term average percent correct on the question, so more difficult questions are farther to the left. We see that all scenarios remained below their long-term average percent correct, which makes sense—most students use NTS to prepare for the final exam, and thus the long-term accuracy is higher due to increased experience with the material. In the post-test, students answered questions with higher correlation to true difficulty (closer to a diagonal line through Figure 4), which is what we would hope to see as students become more familiar with the content. The correlation between the long-term accuracy on scenarios and correct ratio within the assessment is noticeably lower for the pre-test than the post-test, due largely to the scenarios on newer topics (two independent means and mean difference). It is interesting to look at the two scenarios concerning two independent means—one had a major improvement but the other, more difficult, scenario did not. This might suggest that while students were learning and improving, they still had room to grow on the most challenging new-topic scenarios.

When we focus on users (n = 230) of NTS compared to non-users (n = 1569) during the time period between the pre-test and post-test, where users are defined as students who completed at least one session of 10 scenarios, the average improvement for users is significantly larger than for non-users, t(179.99) = 3.30, p = .001. The average score for NTS users was estimated between .234 and .930 points higher compared to non-users.

We also found that more usage of NTS corresponded to greater improvements in scores, $b = .247, t(1024) = 4.16, p < .001$. For each additional session of NTS (measured as 10 completed practice scenarios), the students scored, on average, .25 points higher. Most students who used NTS during this time completed 5 or fewer sessions, so there is likely an unobserved ceiling effect to this benefit.

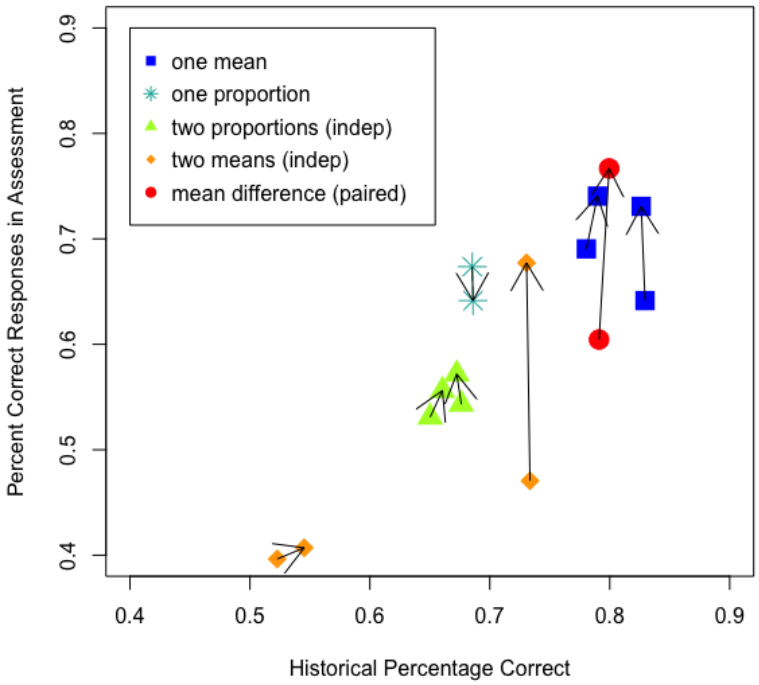


Figure 4.4: Change in average performance for matched scenarios

We explored the possibility that students who used NTS were particularly high-achieving or advanced students by comparing users to non-users on previous exam scores, $t(340.57) = 2.04$, $p = .0424$, and pre-test scores, $t(297.75) = 1.00$, $p = .32$. While there is a small p-value attached to the comparison of exam scores, there is not a practical difference—the difference in population mean exam scores is estimated to be between 0.6549 and 1.7410 percentage points with 95% confidence. Thus it seems that there are not dramatic pre-existing differences in students' performance between the two usage groups.

To further explore how both past performance and NTS usage relate to post-test score, we performed a regression analysis with usage level as a factor. The results are summarized in Table 4.3. Students' Exam 1 performance explained a portion (up to 2 points out of 8) of the final post-test score, the pre-test accounts for a substantial baseline, and amount of NTS use between pre-test and post-test

contributes significantly.

	Estimate	Std Error	t value	Pr(> |t|)
Intercept	2.7711	0.2888	9.592	< 0.0001 ***
Exam 1 score (out of 75)	0.0091	0.0105	0.868	0.3859
Pre-test (out of 8)	0.4451	0.0281	15.851	< 0.0001 ***
1 NTS completed	0.3961	0.2194	1.805	0.07136 ‘
2 NTS completed	0.7707	0.2760	2.792	0.0053 ***
3 NTS completed	0.9229	0.4026	2.292	0.0221 **
>3 NTS completed	0.9915	0.3548	2.795	0.0053 ***

Adj. $R^2 = 0.2128$, F = 47.14 on 6 and 1018 df, p < 0.0001

Table 4.3: Regression results for post-test score (out of 8)

Figure 4.5 shows how students performed on the NTS-related assessments, based on when they used NTS. Here, we can see how use beyond the post-test continued to relate to student performance on the second graded exam. The student groups who used NTS before each assessment performed best on the subsequent assessment. The group in red who used NTS between the pre-test and post-test, but not between the post-test and the second exam, is particularly interesting. This group had the highest average score on the post-test (for which they practiced before), but the lowest average score on the NTS portion of the second exam (for which did not use NTS prior). This suggests that the effects of practice can wear off if the practice is not continued throughout the course. However, Ebbinghaus [22] argued that relearning concepts is easier and faster than learning them for the first time—so this group may be a prime candidate for targeted intervention to encourage continued practice.

4.4 Relevance to Other Chapters in Dissertation

For this chapter, we collected pre- and post-test data from students to evaluate their progress towards mastery of a particular statistics skillset. We made use of paired t-tests to investigate within-person improvements and ordinary least squares regression to model the extent of improvements based on engagement with the NTS

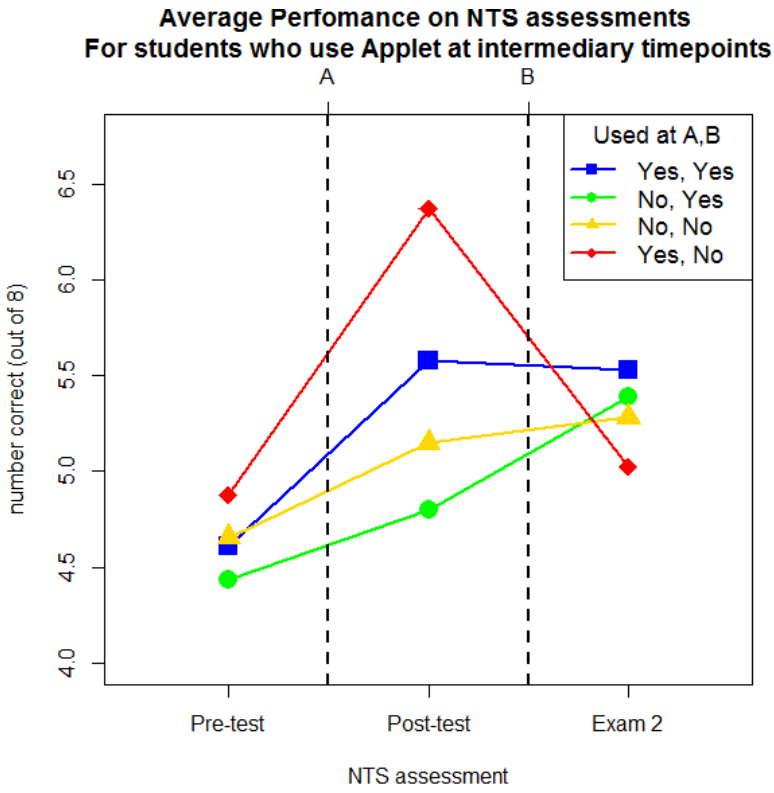


Figure 4.5: Performance based on usage behaviors

applet. These methods facilitate a particular type of assessment: whether (and how much) this learning tool contributes overall to student learning.

However, this practice tool produces a rich dataset on a growing library of questions, which can be used to further explore student mastery over time. By switching our focus from within-person changes to between-person differences, we can further explore the relationship between students and the items in the applet. This allows us to identify features of both the items and the students, which could ultimately lead to a more tailored student learning experience.

The means by which this further exploration is possible is item response theory (IRT). IRT is commonly used in psychometrics to design and analyze items on tests. In an IRT model, the characteristics of each item, such as difficulty, are used to describe the probability that a student with a given ability level θ will answer the item correctly. More precisely, the probability p_i of a item i receiving a correct

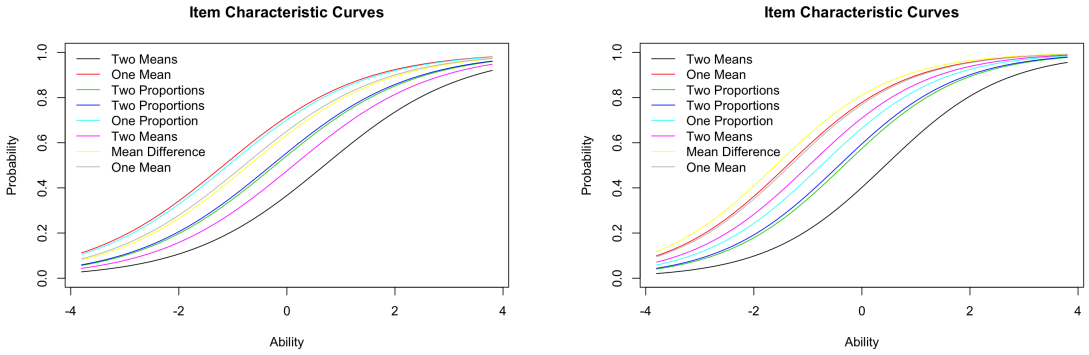
response from a person with ability θ is:

$$p_i = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \tag{4.1}$$

where the item parameters a_i, b_i, and c_i describe the discrimination, difficulty, and chance from guessing, respectively. Thus, each item is described by its own unique function—known as an item characteristic curve (ICC). We can view these ICCs as basis sets, where the parameters have both physical/geometric and practical/content interpretations. One can place constraints on these parameters to test hypotheses about landmarks in the ICCs. The discrimination, a_i, is also the slope of the curve. It describes how small of a difference in ability (plotted on the x-axis) is needed to result in a meaningful difference in the probability of responding correctly. The difficulty, b_i, is the location of the curve. The final parameter, c_i, is an asymptotic minimum which is interpreted as a guessing baseline. For questions like those in the NTS pre-test and post-test with 5 answer choices, a student starts with a probability of 20% of getting the question correct—the probability of selecting the right answer when guessing randomly.

To illustrate how we can model these properties of the individual items, we first focus on only the difficulty of the items based on student performance. Thus we can fit a one-parameter, or Rasch, item response model. Figure 4.6 shows the corresponding ICCs for the 8 items of each assessment.

We can see that the order of the items (they are matched on color) changes between the pre-test and the post-test. There are two possible explanations: the item pairings changed (their true difficulties do not match across assessments) or the students changed (by learning more about some concepts, and maybe forgetting some about other concepts). The item difficulties reported in this paper are treated as true known difficulties, since they are based on tens of thousands of responses. Thus, it seems reasonable to conclude that the changes in the difficulty ratings of the



(a) ICCs for pre-test items

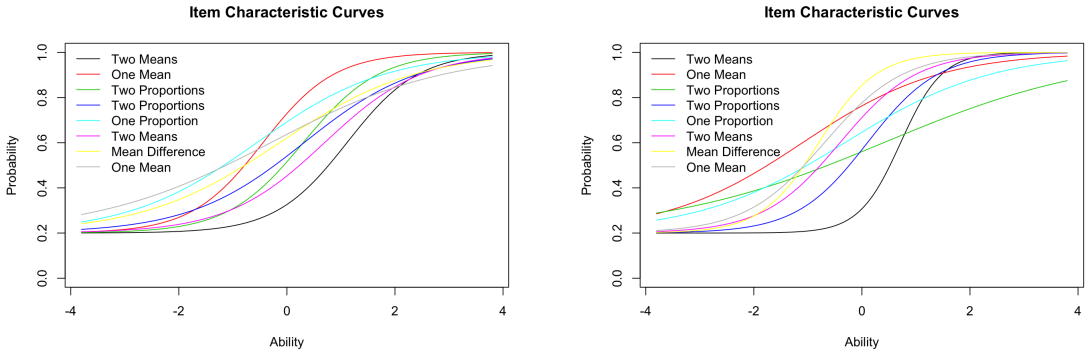
(b) ICCs for post-test items

Figure 4.6: One-parameter models of assessment items

questions reflect the educational progress of the students being evaluated. We can compare the curves from Figure 4.6a and Figure 4.6b and see that the reordering of the items reflects the improvements shown in Figure 4.4. For example, the question on mean differences is rated as the least difficult in the post-test, and this was the question that most students answered correctly on the latter assessment.

This one-parameter model was just one option for fitting the data, and its parameterization allows us to draw conclusions about the parameter being estimated—the difficulty. Just as a change is basis set was shown in Chapter III to have potentially dramatic impacts of the results, a change in basis set to a 3-parameter, or Birnbaum, IRT model may change our findings in this setting. The intuition is that each item has its own 3-parameter logistic function, so it is analogous to having a family of curves parameterized by 3 values that are interpretable in the domain and each item is associated with a particular curve. This allows comparison across curves on the basis of these interpretable parameters, a feature discussed in Chapter III.

Figure 4.7 shows the fitted ICCs for the same pre-test and post-test data, now using the 3-parameter model. The guessing parameter was fixed at 0.2 to reflect the 20% chance of getting the correct answer by guessing alone. The remaining two parameters allow both the discrimination and difficulty of the questions to vary.



(a) ICCs for pre-test items

(b) ICCs for post-test items

Figure 4.7: Three-parameter models of assessment items

Here, we notice that some items have less discrimination (lower slope) than others, and these items are not the same between the pre-test and post-test. For example, the questions concerning two proportions should be matched on difficulty, but in the post-test they appear to have dramatically different discrimination—perhaps these questions are not as similar as we had assumed.

Overall, these results again highlight the potential differences between two basis set options. The two models reflect different assumptions about the data. The one-parameter model assumes that there is no difference in discrimination, while the three-parameter model allows both discrimination and a fixed adjustment for guessing. As with the examples in Chapter III, we need to make use of our existing knowledge about the data to guide us to the correct basis set—otherwise we have two reasonable models with conflicting results.

By thinking about ICCs as basis functions with interpretable parameters, we gain a simple way to visualize the properties of each item in terms of landmarks on the curves. These important landmarks can be tested across groups and over time. Thus, the parameters provide additional explanations for observed differences in the proportion correct by allowing testing across individuals, groups, conditions, and time in terms of the discriminability, difficulty, and guessing rate of each item.

In this way, NTS is also self-evaluating—the gaps in the parameter space of the item characteristic curves will suggest properties of new scenarios to help students practice and keep the applet appropriately challenging for the course. The applet could also be modified to be adaptive for each student, becoming increasingly challenging as the student gains mastery. Similar approaches have been implemented in the field of knowledge tracing [12]. Not only could this prevent students from becoming discouraged by receiving questions outside of their current ability, but it would also allow instructors to see a current estimated mastery level for any student who uses the applet regularly. Thus, the IRT framework viewed as an implementation of basis sets can be used to guide future interventions. We discuss this in more detail in the last chapter of the dissertation.

4.5 Conclusion

We have shown an example of an applet that promotes productive practice of a historically challenging aspect of introductory statistics classes. This applet is generally well-received by students and instructors. We identified features for applet success which complement those outlined by GAISE [27]: feedback for all responses, whether correct or not; summary of performance to users; repeated practice; and portable, on-the-go usability.

On average, students who use the applet to practice identifying the appropriate statistical method show gains on subsequent assessments of this skill. There are other possible explanations of the gains we see. The effect of students naturally improving over time is confounded with NTS use. Historically, improvement over time is common in this course. However, practice is still an important aspect of student learning in any discipline.

Hsu [33] found that instructor-made programs were more effective than commercially available programs. We have found this to be true in our case, and believe that

providing resources such as NTS helps students to adjust to the workload of university courses. Our findings support the conclusion that NTS helps students identify appropriate statistical methods for real-world problems.

CHAPTER V

Conclusion

The aim of this dissertation has been to emphasize the need to select the appropriate representation of data so that it connects with the theoretical models and hypotheses under investigation. By emphasizing this aspect of the model-building process, we can better reflect the underlying science. It also opens the doors to testing theoretical frameworks directly, creating a clear path towards making informed updates to our understanding of the phenomena being examined. The basis set framework proposed and explored in this dissertation makes use of more modern statistical and computing methodology than the methods it has been shown to replace, while retaining these prior methods as special cases.

We began by focusing on a specific applied problem: the analysis of Event-Related Potential (ERP) data. ERP components help researchers understand brain responses to sensory, cognitive, and motor events through changes in the known features of established waveform shapes. Typically, ERP studies focus on amplitude differences across groups or experimental conditions. There are two common summary metrics that are used: the local maximum and the average. Despite our concerns rooted in extreme value theory, we found that the use of local maxima in the standard testing framework is not unreasonable. The distributional assumptions for standard tests will be met when using either the maximum or the average. However, these

two summary metrics can yield contradictory results. They also yield different error rates, depending on the experimental design.

We began our simulation study by specifying a data-generating model that reflects the established features described in existing research. Based on this model, we found cases where each metric excelled. While both summary metrics generated large Type I and Type II error rates when the analysis window failed to capture unmodeled latency variability, the maximum was more robust to small perturbations from the ideal window location and could recover some performance with a wider window. The average outperformed the maximum in studies with unequal numbers of trials, such as those utilizing the oddball paradigm.

Based on our results, we concluded that one summary metric was not universally better than the other. Instead, anticipated features in the data, the costliness of certain errors, or the desired interpretation may need to be stated to justify the choice of summary metric. However, our study was not exhaustive. We made use of a common experimental design and analytic approach, but the high variability in experimental design of ERP studies may prevent these results from being generalizable to all settings. Also, other approaches to ERP analyses have been proposed. Many of these make use of modern computational power and more recent developments in statistics. In the future, we can broaden the scope of study designs and analysis protocols when comparing alternatives to the basis set approach we propose at the end of Chapter II.

The findings of our first paper led to two ideas that were pursued in Chapter III. First, the choices made during seemingly trivial intermediary steps in the modeling procedure can have dramatic impact on the final results of a study. Thus, the choice of basis set is critical. A basis set should make use of known properties of the data, as well as testable hypotheses, by incorporating landmark points in the functional form. Second, the data-generating model used to simulate ERP data can be used

as an analytic model. This is a specific example of a basis set—the height, width, and temporal location of the component being modeled are reflected by parameters that can vary independently or be mathematically constrained. A major advantage of this approach over existing methods is that the processing and analysis is conducted simultaneously, and thus we avoid order effects. We used this basis set to reanalyze an ERP study. While the primary finding stayed the same, by using a basis set we showed additional properties of the dataset. We also demonstrated the use of this basis set in a Regression Spline Mixed Model (RSMM) framework for exploratory analysis of multi-component ERP waveforms.

The basis set methodology we have outlined allows for meaningful landmarks of longitudinal data to be modeled and tested directly. It has broad utility, including most cases where latent growth curves are currently used. The neuroscience application we have described in Chapter III can easily be extended to other psychophysiological measures. It can also be used to explore spatio-temporal relationships within multiple time series in ERP and other contexts. The broad potential for applications allows for continued work in varied domains where unique properties of the time series will require the development and investigation of new basis sets.

An option that we have not explored throughout this work involves allowing a broad library of basis set functions to be fit to the data, rather than designing and fitting a single basis set that facilitates testing of specific hypotheses. This reflects the difference between exploratory data analysis and hypothesis testing. A fitting algorithm could be defined to use principles from machine learning to exclude basis sets that do not fit the data well, and retain a low-dimensional combination of basis sets that combine to provide the best representation of the data. While this option runs the danger of overfitting the data, it could also call attention to use of basis sets that were not otherwise being considered as possible reflections of the science. This could inspire follow-up studies, and possibly updates to our theoretical understand-

ing that would otherwise not have been imagined due to the constraints of existing paradigms.

Because collaborating on scientific research requires communication about the underlying scientific model, it is important to build a common understanding of the properties and appropriate uses for a variety of potential models. Thus, Chapter IV explored an applet designed to help introductory statistics students identify the appropriate statistical model for a given research scenario. While natural improvement over time is confounded with improvement from applet use and students self-selected into use categories rather than being randomly assigned, we did find strong evidence that performance improved significantly for students who used the applet. Not only did we highlight the potential for using technology to help students build this important skill, but we also demonstrated another use case for basis sets. The Item Characteristic Curve (ICC) of each scenario can be reimagined as a basis function with interpretable parameters for the difficulty, discrimination, and guessing probability relative to a student's current ability.

Viewing ICCs in this way allows instructors to focus on specific aspects of each question that might not have been obvious or intentional as they wrote the item. These properties can then be leveraged in powerful ways. For example, a simple computer-adaptive test could provide increasingly difficult questions in each area to students until they begin answering incorrectly, then continue with questions at that level for that concept until an estimated improvement in latent ability suggests a readiness for more difficult questions. Gaps in the test bank may become more apparent by exploring the values of each parameter across each type of question. Even student performance and motivation may be better understood by examining the questions being presented. If the questions are high in discrimination, a small increase in a student's understanding of the topic could lead to a quick gain in performance. However, if the questions are too high in difficulty, presenting easier questions may

help prevent the student from becoming discouraged.

These examples of understanding how the parameters of a model operate and are interpreted, and thus which levers to pull for a successful intervention, are universal to the broader class of models built on meaningful basis sets. Not only can a basis set provide a convenient means for analyzing data in a flexible framework, but it can also be informative to the researcher for guiding next steps—either an intervention, a future study, or even a modification to the hypothesized model.

APPENDIX

APPENDIX A

Appendix for Chapter IV

A.1 Questions from assessments

Quiz	Topic	Question	Past % Correct
Pre-test	Two independent means	A controversial study claims that left-handed students perform better on the SAT than right-handed students. University officials wish to refute it and ask you to provide an estimate for the true difference in SAT scores among left-handed and right-handed applicants to UM. From the population of all recent applicants with SAT scores on file, 1,000 are selected at random and sent an email asking them which is their dominant hand for writing.	0.5227
Post-test	two independent means	Researchers want to look at the effect of taking vitamins on how often children get sick. 100 children were assigned to either take a daily vitamin or take no vitamins for an entire school year. Then their parents reported the number of days of school were missed due to illness that year.	0.5453

Pre-test	one mean	Professors at the University of Michigan believe that students attend 3 hours of classes per day on average. To assess this belief, a random sample of UM students will be selected and asked to report how many hours of classes they attend on a typical class day	0.7805
Post-test	one mean	To test the quality of their pens, a company uses a robot to continually write with a random selection of pens from the batch produced each day. For each pen in the sample, the time until the pen stops writing cleanly (either from lack of ink or a gummed up point) is recorded. The quality guidelines say any batch which lasts less than 15 hours on average may not be shipped. The company brings you the data for their latest batch; can they ship the pens?	0.7898
Pre-test	two proportions	The dean of a college is looking at whether Humanities majors are more satisfied with their choice of major than science and engineering majors. He selects a random sample of students from all majors across the campus, and asks each student whether they are satisfied with their choice of majors. He then asks you to look at his data and provide an estimate for what percentage different the two levels of satisfaction are.	0.6503
Post-test	two proportions	A local cookie business wants to look into how to best preserve their cookies. Their study will involve taking a batch of cookies and wrapping some in standard plastic wrap and some in a new foil package. After being stored overnight, each cookie will be examined to see whether or not it is stale.	0.6605

Pre-test	two proportions	A dental floss company wants to know if adults are more likely to floss daily than children. From their database of customers, they plan to randomly survey families and record respondents ages whether they floss daily. The company considers anyone 18 and older an adult.	0.6762
Post-test	two proportions	A swimming school recently hired a new instructor, and wants to ensure that the new instructor is performing well. A random sample of students from each of the current classes of the new instructor and a seasoned instructor are given a basic swimming test, and the school would like to figure out whether a larger percent of kids passed the test under the seasoned instructor.	0.6724
Pre-test	one proportion	A researcher thinks that more than 80% of squirrels in Ann Arbor are overweight. A research team collects squirrels from around town and weighs them in order to test this hypothesis. If they weigh more than 2.5 pounds, they are considered overweight.	0.6854
Post-test	one proportion	In recent years, the media has bemoaned the low voter turnout to presidential elections. A polling agency wants to know if the next elections are likely to reverse the trend, and ask you to collect a sample to provide an estimate for voter turnout.	0.6860
Pre-test	two independent means	Many stores offer a cheaper store-brand product that corresponds to a more expensive brand-name product. A study is planned to compare a brand-name rice cereal with mini marshmallows versus the corresponding store-brand cereal in terms of the mean number of marshmallows per 12-ounce box. A random sample of 25 boxes of each brand of cereal will be selected and the number of marshmallows in each box of the 50 selected boxes will be counted to conduct this comparison.	0.7335

Post-test	two independent means	Of the male and female law students who have taken an introduction to probability course, a law professor is interested in learning whether there is a significant difference between the average GPAs of each group.	0.7306
Pre-test	mean difference (paired)	A team of psychologists is studying personality differences in males and females. One character trait they are studying is empathy, which is often measured on a scale of 0-100 using a personality test. To control for family effects the study is conducted using sister-brother combinations. To examine whether females more empathetic than males, a random sample of brothers and sisters are given a personality test to determine their empathy scores.	0.7911
Post-test	mean difference (paired)	A researcher wants to know the difference in time spend watching TV for men and women. He asks 50 heterosexual couples to report the amount of time spent watching TV per week, since he suspects that a significant other may influence the amount of time a person watches TV and wants to control for this.	0.7995
Pre-test	one mean	An Introductory Statistics class requires at least a good understanding of basic math. A professor wants to check if her class has the necessary skills. She doesn't want to make all 1,500 students undergo the test, so sixty students in the class are chosen at random and given a short math quiz. If the class average is below 80 on the exam, she plans to hold extra classes to teach the basics. You are given the results from the exam and asked to see if she will need to hold the extra classes.	0.8294
Post-test	one mean	A coach wants to know how many hours athletes exercise per week. A random sample of 100 of his athletes was taken and the number of hours per week each athlete exercised was recorded.	0.8262

Table A.1: Questions from assessments

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Anderson-Cook, C. M. and Dorai-Raj, S. (2003), "Making the Concepts of Power and Sample Size Relevant and Accessible to Students in Introductory Statistics Courses using Applets," *Journal of Statistics Education*, 11.

[2] Balakrishnan, N. (2013), *Handbook of the Logistic Distribution*, Statistics: A Series of Textbooks and Monographs, Taylor & Francis.

[3] Bates, S. P., Galloway, R. K., and McBride, K. L. (2012), "Student-Generated Content: Using PeerWise to Enhance Engagement and Outcomes in Introductory Physics Courses," in *AIP Conference Proceedings, 1413*, pp. 123–126.

[4] Begleiter, H. (1999), "EEG Database Data Set," `https://archive.ics.uci.edu/ml/datasets/EEG+Database`.

[5] Bloom, B. S. (1969), *Taxonomy of Educational Objectives: The Classification of Educational Goals*, Longman Group.

[6] Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012), "The Origin of Extracellular Fields and Currents–EEG, ECoG, LFP and Spikes," *Nature Reviews Neuroscience*, 13, 407–420.

[7] Chance, B., Ben-Zvi, D., Garfield, J., and Medina, E. (2007), "The Role of Technology in Improving Student Learning of Statistics," *Technology Innovations in Statistics Education*, 1.

[8] Chance, B. and Rossman, A. (2006), "Using simulation to Teach and Learn Statistics," in *Proceedings of the Seventh International Conference on Teaching Statistics*, pp. 1–6.

[9] Chew, V. (1968), "Some Useful Alternatives to the Normal Distribution," *The American Statistician*, 22, 22–24.

[10] Cohen, M. X. (2014), *Analyzing Neural Time Series Data*, MIT press.

[11] Cook, E. W. I. and Miller, G. A. (1992), "Digital Filtering: Background and Tutorial for Psychophysiologists," *Psychophysiology*, 29, 350–367.

[12] Corbett, A. T. and Anderson, J. R. (1994), "Knowledge tracing: Modeling the Acquisition of Procedural Knowledge," *User Modeling and User-Adapted Interaction*, 4, 253–278.

[13] Cudeck, R. and du Toit, S. H. C. (2010), "Multivariate Behavioral A Version of Quadratic Regression with Interpretable Parameters," *Multivariate Behavioral Research*, 37, 501–519.

[14] David, H. A. and Nagaraja, H. N. (1970), *Order statistics*, Wiley Online Library.

[15] de Haan, L. (1976), "Sample Extremes: an Elementary Introduction," *Statistica Neerlandica*, 30, 161–172.

[16] Dien, J. (1998), "Issues in the Application of the Average Reference: Review, Critiques, and Recommendations," *Behavior Research Methods, Instruments, & Computers*, 30, 34–43.

[17] Dien, J. and Santuzzi, A. (2004), "Application of Repeated Measures ANOVA to High-Density ERP Datasets: A Review and Tutorial," in *Event Related Potentials: A Methods Handbook*, ed. Handy, T., Cambridge, MA: MIT Pres, chap. 4, pp. 57–82.

[18] Dinov, I. D., Christou, N., and Sanchez, J. (2008), "Central Limit Theorem: New SOCR Applet and Demonstration Activity," *Journal of Statistics Education*, 16, 1–15.

[19] Donchin, E. and Heffley, E. (1978), "Multivariate Analysis of Event-Related Potential Data: A Tutorial Review," in *Multidisciplinary Perspectives in Event-Related Brain Potential Research*, pp. 555–572.

[20] Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., Polich, J., Reinvang, I., and Van Petten, C. (2009), "Event-Related Potentials in Clinical Research: Guidelines for Eliciting, Recording, and Quantifying Mismatch Negativity, P300, and N400," *Clinical Neurophysiology*, 120, 1883–1908.

[21] Dweck, C. S. (2006), *Mindset: The New Psychology of Success*, Random House.

[22] Ebbinghaus, H. (1964), *Memory: A Contribution to Experimental Psychology*, Dover Publications.

[23] Esseen, C. G. (1956), "A Moment Inequality with an Application to the Central Limit Theorem," *Scandinavian Actuarial Journal*, 1956, 160–170.

[24] Estes, W. K. (1956), "The Problem of Inference from Curves Based on Group Data," *Psychological Bulletin*, 53, 134–140.

[25] Fisher, R. A. and Tippett, L. H. C. (1928), "Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample," *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

[26] Friedman, D., Cycowicz, Y. M., and Gaeta, H. (2001), "The Novelty P3: An Event-Related Brain Potential (ERP) Sign of the Brain's Evaluation of Novelty," *Neuroscience and Biobehavioral Reviews*, 25, 355–373.

[27] GAISE College Report ASA Revision Committee (2016), "Guidelines for Assessment and Instruction in Statistics Education College Report," http://www.amstat.org/education/gaise.

[28] Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

[29] Gnedenko, B. (1943), "Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire," *Annals of Mathematics*, 44, 423–453.

[30] Gonzalez, R. (2009), *Data Analysis for Experimental Design*, Guilford Press.

[31] Gonzalez, R. and Wu, G. (1999), "On the Shape of the Probability Weighting Function," *Cognitive psychology*, 38, 129–166.

[32] Helwig, N. E. (2015), *eegkit: Toolkit for Electroencephalography Data*, R package version 1.0-2.

[33] Hsu, Y. (2003), "The Effectiveness of Computer-Assisted Instruction in Statistics Education: A Meta-Analysis," Ph.D. thesis, University of Arizona, Tucson, Retrieved from ProQuest Dissertations and Theses database (UMI No. 3089963).

[34] Johnson, N. L. and Kotz, S. (1970), *Continuous Univariate Distributions*, vol. 2, John Wiley & Sons Inc.

[35] Joyce, C. and Rossion, B. (2005), "The Face-Sensitive N170 and VPP Components Manifest the Same Brain Processes: The Effect of Reference Electrode Site," *Clinical Neurophysiology*, 116, 2613–2631.

[36] Kahle, D. (2014), "Animating Statistics: A New Kind of Applet for Exploring Probability Distributions," *Journal of Statistics Education*, 22, 1–21.

[37] Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., Luu, P., Miller, G. A., and Yee, C. M. (2014), "Committee Report: Publication Guidelines and Recommendations for Studies Using Electroencephalography and Magnetoencephalography," *Psychophysiology*, 51, 1–21.

[38] Kutas, M. and Hillyard, S. A. (1980), "Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity," *Science*, 207, 203–205.

[39] Lindstrom, M. J. and Bates, D. M. (1990), "Nonlinear Mixed Effects Models for Repeated Measures Data," *Biometrics*, 46, 673–687.

[40] Lovett, M. and Greenhouse, J. (2000), "Applying Cognitive Theory to Statistics Instruction," *The American Statistician*, 54, 196.

[41] Luck, S. J. (2005), "Ten Simple Rules for Designing ERP Experiments," in *Event-Related Potentials: A Methods Handbook*, ed. Handy, T., The MIT Press, pp. 17–32.

[42] — (2014), *An Introduction to the Event-Related Potential Technique*, MIT press.

[43] Mackenzie, M. L., Donovan, C., and McArdle, B. (2005), "Regression Spline Mixed Models: A Forestry Example," *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 394–410.

[44] Mallot, H. A. (2013), *Computational Neuroscience: A First Course*, vol. 2, Springer.

[45] Michalewski, H., Prasher, D., and Starr, A. (1986), "Latency Variability and Temporal Interrelationships of the Auditory Event-Related Potentials (N1, P2, N2, and P3) in Normal Subjects," *Electroencephalography and Clinical Neurophysiology*, 65, 57–71.

[46] Murray, M. M., Brunet, D., and Michel, C. M. (2008), "Topographic ERP Analyses: A Step-by-Step Tutorial Review," *Brain Topography*, 20, 249–264.

[47] Nurnberger, G. (1989), *Approximation by Spline Functions*, Springer-Verlag.

[48] Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., and Taylor, M. J. (2000), "Guidelines for Using Human Event-Related Potentials to Study Cognition: Recording Standards and Publication Criteria," *Psychophysiology*, 37, 127–152.

[49] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015), *nlme: Linear and Nonlinear Mixed Effects Models*, R package version 3.1-120.

[50] Preacher, K., Hancock, G., Harring, J., and Hancock, G. (2012), "On Interpretable Reparameterizations of Linear and Nonlinear Latent Growth Curve Models," *Advances in Longitudinal Methods in the Social and Behavioral Sciences*, 25–58.

[51] Proakis, J. and Manolakis, D. (1992), *Digital Signal Processing : Principles, Algorithms, and Applications*, Macmillan Publishing Company.

[52] Ramsay, J. O. (2006), *Functional Data Analysis*, Wiley Online Library.

[53] Rice, J. A. and Wu, C. O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves." *Biometrics*, 57, 253–259.

[54] Sandholtz, J. H., Ringstaff, C., and Dwyer, D. C. (1997), *Teaching with Technology: Creating Student-Centered Classrooms*, Teachers College Press.

[55] Schenker, J. D. (2007), "The Effectiveness of Technology use in Statistics Instruction in Higher Education: A Meta-Analysis Using Hierarchical Linear Modelling," Ph.D. thesis, Kent State University, Retrieved from ProQuest Dissertations and Theses database (AAT 328685).

[56] Schneiter, K. and Symanzik, J. (2013), "An Applet for the Investigation of Simpson's Paradox," *Journal of Statistics Education*, 21, 1–20.

[57] Shevtsova, I. (2011), "On the Absolute Constants in the Berry-Esseen Type Inequalities for Identically Distributed Summands," *arXiv preprint arXiv:1111.6554*.

[58] Smith, A. F. M. and Cook, D. G. (1980), "Straight Lines with a Change-Point : A Bayesian Analysis of Some Renal Transplant Data," *Applied Statistics*, 29, 180–189.

[59] Sosa, G. W., Berger, D. E., Saw, A. T., and Mary, J. C. (2011), "Effectiveness of Computer-Assisted Instruction in Statistics: A Meta-Analysis," *Review of Educational Research*, 81, 97–128.

[60] Stuart, A. and Ord, K. (1994), *Kendall's Advanced Theory of Statistics*, vol. 1, New York, NY: Halsted Press, 6th ed.

[61] Svinicki, M. D. (1999), "New Directions in Learning and Motivation," *New Directions for Teaching and Learning*, 80, 5–27.

[62] West, R. W. and Ogden, R. T. (1998), "Interactive Demonstrations for Statistics Education on the World Wide Web," *Journal of Statistics Education*, 6, 1–9.

[63] Zucker, T. A., Moody, A. K., and McKenna, M. (2009), "What Forty Years of Research Says About the Impact of Technology on Learning A Second-Order Meta-Analysis and Validation Study," *Journal of Educational Computing Research*, 40, 47–87.